Characterizing the Etiology of Recurrent Tuberculosis Using Whole Genome Sequencing: Alaska, 2008–2020

Yuri P. Springer[1], **Megan L. Tompkins**[2], **Katherine Newell**[2,3], **Martin Jones**[2,4], **Scott Burns**[1], **Bruce Chandler**[2], **Lauren S. Cowan**[1], **J. Steve Kammerer**[1], **James E. Posey**[1], **Kala M. Raz**[1], **Michelle Rothoff**[2], **Benjamin J. Silk**[1], **Yvette L. Vergnetti**[2], **Joseph B. McLaughlin**[2], **Sarah Talarico**[1]

[1]Division of Tuberculosis Elimination, National Center for HIV, Viral Hepatitis, STD, and TB Prevention, Centers for Disease Control and Prevention, Atlanta, Georgia

[2]Section of Epidemiology, Alaska Division of Public Health, Anchorage, Alaska

[3]Epidemic Intelligence Service, Division of Workforce Development, National Center for State, Tribal, Local, and Territorial Public Health Infrastructure and Workforce, Centers for Disease Control and Prevention, Atlanta, Georgia

[4]Public Health Associate Program, Division of Workforce Development, National Center for State, Tribal, Local, and Territorial Public Health Infrastructure and Workforce, Centers for Disease Control and Prevention, Atlanta, Georgia

Abstract

Background.—Understanding the etiology of recurrent tuberculosis (rTB) is important for effective tuberculosis control. Prior to the advent of whole genome sequencing (WGS), attributing rTB to relapse or reinfection using genetic information was complicated by the limited resolution of conventional genotyping methods.

Methods.—We applied a systematic method of evaluating whole genome single-nucleotide polymorphism (wgSNP) distances and results of phylogenetic analyses to characterize the etiology of rTB in American Indian and Alaska Native (AIAN) persons in Alaska during 2008 to 2020. We contextualized our findings through descriptive analyses of surveillance data and results of a literature search for investigations that characterized rTB etiology using WGS.

Correspondence: Yuri P. Springer, PhD, Division of Tuberculosis Elimination, US Centers for Disease Control and Prevention, 1600 Clifton Rd NE, Atlanta, GA 30329 (vkh3@cdc.gov).

Author contributions. Y. P. S. and B. J. S. conceived the investigation. Y. P. S. and S. T. led the analyses. All authors critically reviewed findings, contributed to the writing, and read and approved the final manuscript.

Potential conflicts of interest. All authors: No reported conflicts.

All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

Disclaimer. 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.

Supplementary Data

Supplementary materials are available at *The Journal of Infectious Diseases* online (http://jid.oxfordjournals.org/). Supplementary materials consist of data provided by the author that are published to benefit the reader. The posted materials are not copyedited. The contents of all supplementary data are the sole responsibility of the authors. Questions or messages regarding errors should be addressed to the author.

Results.—The percentage of tuberculosis cases in AIAN persons in Alaska classified as recurrent episodes (11.8%) was 3 times the national percentage (3.9%). Of 38 recurrent episodes included in genetic analyses, we attributed 25 (65.8%) to reinfection based on wgSNP distances and phylogenetic analyses; this proportion was the highest among 16 published point estimates identified through the literature search. By comparison, we attributed 11 (28.9%) and 6 (15.8%) recurrent episodes to reinfection based on wgSNP distances alone and on conventional genotyping methods, respectively.

Conclusions.—WGS and attribution criteria involving genetic distances and patterns of relatedness can provide an effective means of elucidating rTB etiology. Our findings indicate that rTB occurs at high proportions among AIAN persons in Alaska and is frequently attributable to reinfection, reinforcing the importance of active surveillance and control measures to limit the spread of tuberculosis disease in Alaskan AIAN communities.

Keywords

American Indian or Alaska native (AIAN) persons; etiology; phylogenetic analysis; recurrent tuberculosis; whole genome sequencing

Recurrent tuberculosis disease (rTB) involves a second (recurrent) case of tuberculosis (TB) in a person after a period of treatment and presumptive recovery following an initial case. In the United States, the annual proportion of TB cases that are recurrent cases has been estimated as approximately 5% since at least 1985 [1, 2]; globally, investigations of rTB in areas associated with high burden or involving populations at high risk have reported recurrent proportions as high as 47% [3]. Because recurrent cases can account for a sizable fraction of all TB cases and are associated with poor treatment outcomes and high mortality rates relative to initial cases [1, 4], elucidating the epidemiology of rTB is critical to improving the effectiveness of TB control programs and reducing TB-associated morbidity and mortality.

A recurrent episode can be attributed to either endogenous relapse following reactivation of the *Mycobacterium tuberculosis* (MTB) strain responsible for the initial case or exogenous reinfection involving either a different MTB strain or the same strain responsible for the initial case. Comparing the genotypes of the MTB strain or strains associated with initial and recurrent cases can sometimes help to distinguish between these etiologies. While a finding of genetically divergent genotypes is consistent with reinfection, interpreting genetically similar or indistinguishable genotypes as evidence of relapse may be complicated by multiple factors. Historically, the limited discriminatory power of conventional genotyping methods, such as spoligotyping [5] and MIRU-VNTR strain typing (mycobacterial interspersed repetitive unit–variable number of tandem repeats), was among these [6]. Use of whole genome sequencing (WGS) in recent years has largely overcome this methodological constraint [7]. Whereas spoligotyping and MIRU-VNTR together cover <1% of the MTB genome, WGS typically covers approximately 90%, allowing for more comprehensive discrimination of relapse vs reinfection rTB etiologies [8].

Within the United States, TB incidence and frequencies of select TB risk factors are particularly high among persons who identify with American Indian or Alaska Native

(AIAN) race (hereafter, AIAN persons) [9]. These disparities are especially pronounced in Alaska. During 2010 to 2020, annual age-adjusted TB incidence among AIAN persons in Alaska (mean, 37.7 cases per 100 000 persons) was on average 21 times that among AIAN persons in states other than Alaska (mean, 2.0 per 100 000 persons) and 84 times that among non-Hispanic White persons nationally (mean, 0.5 per 100 000 persons) [10]. Among persons aged ≥15 years, TB cases in AIAN persons in Alaska were 2.7 and 3.9 times more likely to be attributed to recent transmission (defined using the plausible source case method [11]) than cases in AIAN persons in states other than Alaska and in non-Hispanic White persons nationally, respectively; among persons of all ages, the proportions of TB cases that were recurrent episodes were 2.4 and 2.9 times higher [10]. These findings suggest that ongoing TB transmission and rTB are public health issues of particular concern among AIAN persons in Alaska.

We characterized the frequency and etiology of rTB in AIAN persons in Alaska during 2008 to 2020. Using descriptive analyses, we compared patterns of rTB among race/ethnicity groups and contrasted frequencies in Alaska with those across the United States. We then applied a systematic method of evaluating whole genome single-nucleotide polymorphism (wgSNP) distances and results of phylogenetic analyses involving potential source cases to attribute recurrent episodes to relapse or reinfection. We compared these attributions with those that would have been made based on wgSNP distances alone and on results of conventional genotyping methods. To contextualize our findings, we compared our estimated reinfection proportions with published estimates produced by other investigations that characterized rTB etiology using WGS.

METHODS

Descriptive analyses were based on data reported to the Centers for Disease Control and Prevention’s National Tuberculosis Surveillance System. We included all incident cases that met the TB case definition [12] during 1 January 2008 to 31 December 2020, were counted within any of the 50 states or the District of Columbia, and occurred in a person whose self-reported origin of birth was the United States and whose self-reported country of birth was the United States (or missing); note that we use “case” to refer to an administratively distinct instance of disease in a given person, not to a person experiencing disease. We defined a recurrent episode as a TB case in a given person after having been discharged or lost to supervision for >12 consecutive months following a previous case. We assigned cases to 1 of 8 race/ethnicity groups using self-reported race and ethnicity data. For each group, we enumerated the total number of TB cases and the number and percentage of those classified as recurrent cases separately for cases counted in Alaska and the United States. We calculated 95% confidence intervals (CIs) around percentages of TB cases classified as recurrent cases using the Wilson score interval method [13]. We calculated TB incidence using population estimates from the US Census Bureau [14]. Genetic analyses used data collected by the Centers for Disease Control and Prevention’s National Tuberculosis Genotyping Service [15]. In 2004, the service began conventional genotyping of at least 1 MTB isolate from each culture-confirmed TB case in the United States; spoligotyping [5] and MIRU-VNTR strain typing [6] were performed until 2022. In 2018, the National Tuberculosis Molecular Surveillance Center began using WGS for universal genotyping.

WGS data can be used to measure the wgSNP distance between pairs of MTB isolates and for phylogenetic analyses to characterize genetic relatedness among groups of isolates. For investigative purposes, MTB isolates for select cases have undergone WGS (case counted prior to 2018) or WGS and conventional genotyping (prior to 2004) retrospectively.

We focused genetic analyses on recurrent episodes in AIAN persons residing in Alaska with a recurrent case count date during 2008 to 2020. Alaska Department of Health staff identified recurrent cases using medical charts to match each recurrent case with its corresponding initial case. We identified the subset of episodes for which WGS data were available for an initial and recurrent case isolate. One additional recurrent episode counted in 2000 was included because it occurred in a person who had multiple recurrent episodes during 2008 to 2020. Associated cases were counted in 1 of 4 geographic areas: the Aleutians West, Bethel, and Nome Census Areas, as well as the Municipality of Anchorage. To identify potential source cases for these recurrent episodes, we considered the subset of all TB cases counted in Alaska during or prior to 2020 for which WGS data were available for an associated isolate. We performed additional WGS as funding and isolate availability allowed to increase the number of cases for which WGS data were available; cases counted in 1 of the aforementioned geographic areas and with the same GENType (unique genotype based on results of spoligotyping and MIRU-VNTR typing [15]) as 1 or more of the recurrent cases were prioritized.

WGS and wgSNP comparisons were performed as previously described [16] using BioNumerics 7.6.3 (Applied Maths). The median average coverage depth was 74× (range, 20× −234×). A single-nucleotide polymorphism was retained in the wgSNP comparison if coverage among all samples was ≥5 reads (≥1 forward and reverse), it contained no ambiguous/unreliable bases or gaps, and it was ≥12 base pairs from any other single-nucleotide polymorphism. We first performed an individual wgSNP comparison for each recurrent episode to estimate the wgSNP distance between the initial and recurrent case isolates. We then performed an aggregate wgSNP comparison of all cases (including recurrent episode cases) for which WGS data were available for an associated isolate to estimate the pairwise wgSNP distance between each pair of isolates; from this, we identified all clusters of ≥2 isolates with pairwise wgSNP distances ≤10. For each cluster that had at least 1 isolate associated with a recurrent episode, we performed a cluster wgSNP comparison and constructed a phylogenetic tree using the neighbor-joining method and most recent common ancestor placement based on rooting with MTB H37Rv as the outgroup (Figure 1). For a given recurrent case, potential source cases were those that met 3 criteria: (1) counted during the interepisode interval (period between the initial and recurrent case count dates, the latter extended 90 days to account for reporting delays); (2) associated isolate in the same cluster as the recurrent case isolate; (3) pairwise wgSNP distance from the recurrent case isolate less than or equal to the wgSNP distance between the recurrent and corresponding initial case isolates (both distances based on the cluster wgSNP comparison). For each recurrent episode, we estimated potential source case coverage as the proportion of cases (excluding the initial case) counted in the same geographic area as the recurrent case during the interepisode interval and during the period spanning 2 years prior to and 90 days following the recurrent case count date for which WGS data were available for an associated isolate.

Attribution of recurrent episodes to relapse or reinfection based on wgSNP distances and phylogenetic analyses was performed by applying criteria detailed in Table 1. We compared these attributions with those that would have been made based on wgSNP distances alone (criterion 1) and based on conventional genotyping methods. For the latter, we attributed a recurrent episode to reinfection if initial and recurrent case isolates differed at >1 position across the spoligotype octal code and MIRU-VNTR pattern (12- or 24-locus depending on available data) and to relapse if isolates differed at ≤1 position [2]. We calculated 95% CIs around relapse and reinfection proportions using the Wilson score interval method [13].

To contextualize our findings, we summarized results of other investigations of rTB etiology using WGS. We identified relevant publications through peer-reviewed literature searches using PubMed (title/abstract words) and Google Scholar (title word) and the keywords “tuberculosis whole genome” combined with “recurrent,” “recurrence,” “relapse,” or “reinfection.” For the relevant investigations identified, we summarized information on geographic locations of recurrent episodes, genotyping methods used, number of episodes genotyped, duration of interepisode intervals, and point estimates of reinfection proportions.

RESULTS

Among 134 234 incident TB cases reported to the National Tuberculosis Surveillance System during 2008 to 2020, 42 822 (31.9%) were counted within 1 of the 50 states or District of Columbia and in a US-born person (Table 2). Of these, 1663 (3.9%; 95% CI, 3.7%–4.1%) were recurrent cases. Nationally, AIAN persons were associated with the highest percentage (7.0%; 95% CI, 5.9%–8.4%) of TB cases classified as recurrent cases of any race/ethnicity group; the annual percentage for this group ranged from 1.2% (2020; 95% CI, .2%–6.3%) to 13.6% (2014; 95% CI, 8.8%–20.5%; Supplementary Table 1). The percentage for AIAN persons in Alaska (11.8%; 95% CI, 9.4%–14.8%) was 5.6, 3.1, and 2.6 times higher than for non-Hispanic single-race Asian, White, and Black persons nationally, respectively.

Of 66 recurrent episodes in AIAN persons counted in Alaska during 2008 to 2020, 37 (56.0%) had WGS data available for an initial and recurrent case isolate; adding the 1 additional recurrent episode counted in 2000 brought the total number of recurrent episodes included in our genetic analyses to 38. These recurrent episodes involved 29 unique persons: 22 who had TB twice, 5 who had TB 3 times, and 2 who had TB 4 times. Among these persons, 76.3% were male; the median age at recurrence was 52 years (range, 26–77); and the median interepisode interval was 72 months (range, 18–271; Supplementary Table 2). Of the 38 recurrent episodes, 11 (28.9%) were counted in the Aleutians West Census Area, 13 (34.2%) in the Bethel Census Area, 10 (26.3%) in the Nome Census Area, and 4 (10.5%) in the Municipality of Anchorage (Supplementary Table 3). Of 1779 TB cases counted in Alaska during or prior to 2020, 448 (25.2%) had WGS data available for an associated isolate. The aggregate wgSNP comparison identified 8 clusters of ≥2 isolates with pairwise wgSNP distances ≤10 that included at least 1 isolate associated with a recurrent episode. The median potential source case coverage was 70.5% (range, 5%–100%) during the interepisode interval and 77.0% (range, 6.1%–100%) during the 2 years prior to the recurrent case count date.

Among the 38 recurrent episodes, we attributed 25 (65.8%; 95% CI, 49.9%–78.8%) to reinfection, 2 (5.3%; 95% CI, 1.5%–17.3%) to relapse, and 11 (28.9%; 95% CI, 17.0%–44.8%) to indeterminant etiology based on wgSNP distances and phylogenetic analyses (Table 3, Supplementary Tables 2 and 3). The estimated reinfection proportion was highest when attributions were based on wgSNP distances and phylogenetic analyses (65.8%; 95% CI, 49.9%–78.8%), intermediate when based on wgSNP distances alone (28.9%; 95% CI, 17.0%–44.8%), and lowest when based on conventional genotyping methods (15.8%; 95% CI, 7.4%–30.4%). Of the 25 episodes attributed to reinfection based on wgSNP distances and phylogenetic analyses, 14 (56.0%; 95% CI, 37.1%–73.3%) and 19 (76.0%; 95% CI, 56.6%–88.5%) were attributed to relapse based on wgSNP distances alone and on conventional genotyping methods, respectively.

The literature search identified 63 publications, of which 16 were relevant (ie, investigations that characterized rTB etiology using WGS) [17–32]. Among 8 (50.0%) publications that reported point estimates of reinfection proportions (or findings that could be used to enumerate them) based on both wgSNP distances alone and on conventional genotyping methods, reinfection proportions based on the former were higher in 3 (37.5%; Figure 2, Supplementary Table 4). Among 2 (12.5%) publications that reported point estimates of reinfection proportions based on wgSNP distances alone and on wgSNP distances and phylogenetic analyses, reinfection proportions based on the former were lower in both (100%). The remaining 6 (37.5%) publications reported point estimates of reinfection proportions based only on wgSNP distances alone. Overall, reinfection proportions ranged from 0% to 47.5%.

DISCUSSION

Among TB cases reported in US-born persons during 2008 to 2020, the percentage in AIAN persons in Alaska classified as recurrent cases was 3 times the national percentage. Although AIAN persons in Alaska represented only 0.05% of the US population during this period (annual average, excluding 2008) [14], they accounted for 1.3% of TB cases and 4.0% of recurrent cases nationally. Another recent investigation of TB among US-born AIAN persons (2010–2020) found that the prevalence of rTB among AIAN persons in Alaska was 2.4 and 2.9 times higher when compared with AIAN persons in states other than Alaska and non-Hispanic White persons nationally, respectively [10]. These findings demonstrate that rTB among AIAN persons in Alaska is a serious public health concern, and they reinforce the importance of active surveillance and control measures to limit the spread of TB disease.

We estimated that 25 (65.8%; 95% CI, 49.9%–78.8%) recurrent episodes in AIAN persons counted in Alaska during 2008 to 2020 were attributable to reinfection based on wgSNP distances and phylogenetic analyses. While comparing point estimates of reinfection proportions among investigations should be done with caution because estimates can be strongly influenced by setting, focal population, genotyping method, interepisode interval durations, and sample size, this estimate appears to be the highest of any published to date based on an investigation that characterized rTB etiology using WGS (Figure 2, Supplementary Table 4 [33]). Our finding of a high reinfection proportion is consistent with 2 complementary conclusions, both of which align with results of the aforementioned investigation of TB among US-born AIAN persons during 2010 to 2020 [10]. First, rTB among AIAN persons in Alaska attributable to relapse appears to be relatively rare, suggesting the availability and utilization of quality clinical care for TB disease; we previously found that AIAN persons with TB in Alaska were significantly more likely to receive TB treatment partially or completely as directly observed therapy and that TB treatment was completed at significantly higher rates as compared with AIAN persons with TB in states other than Alaska and with non-Hispanic White persons with TB nationally [10]. Second, the risk of rTB among AIAN persons in Alaska may be high due to ongoing MTB transmission in associated communities; we previously found that among persons aged ≥15 years, the prevalence of TB cases in AIAN persons in Alaska attributed to recent transmission [11] was 2.7 and 3.9 times higher, respectively, than in persons in the same 2 comparison groups [10].

Multiple characteristics of our study population were advantageous for a proof-of-concept investigation into the application of WGS and phylogenetic analyses to characterize rTB etiology. First, because the frequency of rTB in this population is high, our findings could inform specific public health actions to reduce rTB and overall TB burden. Second, a high proportion of TB cases in AIAN persons in Alaska, including 34 (89.5%) recurrent episodes, are in persons who reside in small, remote communities. When cases or outbreaks of TB disease occur in these settings, insights into sources of infection provided by contact investigations can be limited because most community members know and interact regularly with one another, complicating identification of meaningful epidemiologic links [34]. Third, because MTB genetic diversity is relatively low in Alaska as compared with other states, insights into rTB etiology provided by conventional genotyping methods were likely limited. During 2008 to 2020, Alaska was associated with 23.8 unique MTB GENTypes per 100 genotyped TB cases in US-born persons, the lowest frequency of any state (Supplementary Table 5). Finally, while use of WGS allowed rTB etiology to be characterized more comprehensively than conventional genotyping methods, we ultimately found that even wgSNP distances alone were often misleadingly small due to low MTB genetic diversity. Among 17 recurrent episodes with a wgSNP distance of 2 to 9 single-nucleotide polymorphisms and attributed to relapse based on wgSNP distances alone, 14 (82.4%; 95% CI, 59.0%–93.8%) were attributed to reinfection based on wgSNP distances and phylogenetic analyses. Our results demonstrate that even in settings involving small populations and low MTB genetic diversity, WGS and attribution criteria based on wgSNP distances and phylogenetic analyses provide an effective means of elucidating rTB etiology.

The literature search results were not entirely consistent with the notion that WGS should have higher power to discriminate rTB etiologies than conventional genotyping methods. Among 9 investigations (including ours) that evaluated rTB etiology based on both wgSNP distances alone and on conventional genotyping methods, point estimates of reinfection proportions based on the former were higher in 4 by an average of 10.9 percentage points (range, 6.0–15.7) and lower in 4 by an average of 4.6 percentage points (range, 0.9–8.7). These results suggest that the discriminatory power of different genotyping methods may be more nuanced and sensitive to factors such as the specific conventional methods used and genotyping completeness (eg, 12- vs 24-locus MIRU-VNTR typing). Among 3 investigations (including ours) that enumerated point estimates of reinfection proportions based on both wgSNP distances alone and on wgSNP distances and phylogenetic analyses, reinfection proportions based on the latter were higher in all 3 by an average of 28.0 percentage points (range, 22.2–36.9). Folkvardsen et al [21] constructed median joining networks and considered network linkage distances and intermediate nodes when making etiologic attributions; Sadovska et al [30] considered the evolution of unique single-nucleotide variants present in recurrent but not initial case isolates. Results of these investigations highlight the value of considering information on genetic distances and broader patterns of relatedness when rTB etiology is characterized using WGS.

Our investigation had several limitations. Of 66 recurrent episodes in AIAN persons counted in Alaska during 2008 to 2020, 29 (43.9%) did not have WGS data available for an initial and recurrent case isolate and were excluded from our genetic analyses. Often this was a result of the initial case occurring so far in the past that an associated isolate was unavailable. Among these 29 recurrent episodes, the median interepisode interval was 406 months (range, 23–744); on average, this was 5.6 times longer than the 38 recurrent episodes included in our genetic analysis. This result suggests that our estimated reinfection proportions might have been higher had we been able to include additional recurrent episodes since relapse after such long periods is rare [35]. While only 12-locus (vs 24-locus) MIRU-VNTR data were available for 13 (34.2%) of our recurrent episodes, more complete conventional genotyping data might have increased but could not have decreased our estimated reinfection proportions. Potential source case coverage for 11 (28.9%) recurrent episodes was <70% (range, 6.1%–66.7%); low potential source case coverage could result in misattribution of a recurrent episode to relapse if the source case was not included in the analyses. Of the 2 recurrent episodes attributed to relapse based on wgSNP distances and phylogenetic analyses (criterion 2), potential source case coverage was 88.0% and 63.6% overall and 100% among cases with the same GENType as the initial and/or recurrent case isolates (ie, the most likely potential source cases for the recurrent case). The 1 recurrent episode associated with indeterminant etiology based on low potential source case coverage (criterion 7) was associated with a value of 6.1%; meaningfully increasing coverage for this recurrent episode was logistically and financially infeasible. Finally, we did not evaluate our attributions against clinical data on treatment success or failure sometimes available in medical records.

We documented a high percentage of TB cases among AIAN persons in Alaska during 2008 to 2020 that were recurrent cases. Despite the small population sizes and low MTB genetic diversity associated with the communities in which most of our focal recurrent episodes occurred, our investigation demonstrates how WGS attribution criteria involving both genetic distances and patterns of relatedness can distinguish rTB etiologies more reliably and accurately than other genetic methods. Our finding of a high reinfection proportion is consistent with the elevated TB incidence and ongoing TB transmission in Alaska and reinforces the importance of intensive public health surveillance and robust TB control programs to limit MTB transmission in Alaskan AIAN communities.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments.

The authors thank all public health staff who contribute data to the National Tuberculosis Surveillance System. This investigation was determined not to constitute human subjects research by the Centers for Disease Control and Prevention and did not require approval by an institutional review board because data were collected and analyzed as part of routine public health surveillance. The demographic and gene sequence data used in this investigation were collected and analyzed for the purposes of public health surveillance and are routinely available only to state and local public health department jurisdictions due to necessary patient privacy protections. Requests to access these data should be directed to Sarah Talarico, mzi4@cdc.gov.

Financial support.

This work was funded by the Centers for Disease Control and Prevention; excluding staff hours, no funds were allocated specifically for this investigation.

References

1. Kim L, Moonan PK, Yelk Woodruff RS, Kammerer JS, Haddad MB. Epidemiology of recurrent tuberculosis in the United States, 1993–2010. Int J Tuberc Lung D 2013; 17:357–60.
2. Interrante JD, Haddad MB, Kim L, Gandhi NR. Exogenous reinfection as a cause of late recurrent tuberculosis in the United States. Ann Am Thorac Soc 2015; 12:1619–26. [PubMed: 26325356]
3. Mirsaeidi M, Sadikot RT. Patients at high risk of tuberculosis recurrence. Int J Mycobacteriol 2018; 7:1–6. [PubMed: 29516879]
4. Panjabi R, Comstock G, Golub J. Recurrent tuberculosis and its risk factors: adequately treated patients are still at high risk. Int J Tuberc Lung Dis 2007; 11:828–37. [PubMed: 17705947]
5. Kamerbeek J, Schouls L, Kolk A, et al. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. J Clin Microbiol 1997; 35:907–14. [PubMed: 9157152]
6. Supply P, Allix C, Lesjean S, et al. Proposal for standardization of optimized mycobacterial interspersed repetitive unit–variable-number tandem repeat typing of *Mycobacterium tuberculosis*. J Clin Microbiol 2006; 44:4498–510. [PubMed: 17005759]
7. Kizny Gordon A, Marais B, Walker TM, Sintchenko V. Clinical and public health utility of *Mycobacterium tuberculosis* whole genome sequencing. Int J Infect Dis 2021; 113(Suppl 1):S40–2. [PubMed: 33716192]
8. Merker M, Kohl TA, Niemann S, Supply P. The evolution of strain typing in the *Mycobacterium tuberculosis* complex. Adv Exp Med Biol 2017; 1019:43–78. [PubMed: 29116629]
9. Springer YP, Kammerer JS, Silk BJ, Langer AJ. Tuberculosis in indigenous persons—United States, 2009–2019. J Racial Ethn Health Disparities 2021; 8:1–6. [PubMed: 33104967]
10. Springer YP, Kammerer JS, Felix D, et al. Using geographic disaggregation to compare tuberculosis epidemiology among American Indian and Alaska Native persons—USA, 2010–2020. J Racial Ethn Health Disparities 2024. doi:10.1007/s40615-024-01919-z
11. France AM, Grant J, Kammerer JS, Navin TR. A field-validated approach using surveillance and genotyping data to estimate tuberculosis attributable to recent transmission in the United States. Am J Epidemiol 2015; 182: 799–807. [PubMed: 26464470]
12. Centers for Disease Control and Prevention. Reported tuberculosis in the United States. Available at: https://www.cdc.gov/tb/statistics/reports/2021/default.htm. Accessed 3 May 2023.
13. Wilson EB. Probable inference, the law of succession, and statistical inference. J Am Stat Assoc 1927; 22:209–12.
14. US Census Bureau. American Community Survey (ACS) 5-year public use microdata samples (PUMS) dataset. Available at: https://data.census.gov/mdat/#/. Accessed on 3 May 2023.
15. Ghosh S, Moonan PK, Cowan L, Grant J, Kammerer JS, Navin TR. Tuberculosis genotyping information management system: enhancing tuberculosis surveillance in the United States. Infect Genet Evol 2012; 12:782–8. [PubMed: 22044522]

16. Nelson KN, Talarico S, Poonja S, et al. Mutation of *Mycobacterium tuberculosis* and implications for using whole-genome sequencing for investigating recent tuberculosis transmission. Front Pub Health 2022; 9:790544. [PubMed: 35096744]

17. Asare P, Osei-Wusu S, Baddoo NA, et al. Genomic epidemiological analysis identifies high relapse among individuals with recurring tuberculosis and provides evidence of recent household-related transmission of tuberculosis in Ghana. Int J Infect Dis 2021; 106:13–22. [PubMed: 33667696]

18. Bryant JM, Harris SR, Parkhill J, et al. Whole-genome sequencing to establish relapse or re-infection with *Mycobacterium tuberculosis*: a retrospective observational study. Lancet Resp Med 2013; 1:786–92.

19. Dippenaar A, Vos D, Marx M, et al. Whole genome sequencing provides additional insights into recurrent tuberculosis classified as endogenous reactivation by IS6110 DNA fingerprinting. Infect Genet Evol 2019; 75:103948. [PubMed: 31276801]

20. Du J, Li Q, Liu M, et al. Distinguishing relapse from reinfection with whole-genome sequencing in recurrent pulmonary tuberculosis: a retrospective cohort study in Beijing, China. Front Microbiol 2021; 12:754352. [PubMed: 34956119]

21. Folkvardsen DB, Norman A, Rasmussen EM, Lillebaek T, Jelsbak L, Andersen ÅB. Recurrent tuberculosis in patients infected with the predominant *Mycobacterium tuberculosis* outbreak strain in Denmark: new insights gained through whole genome sequencing. Infect Genet Evol 2020; 80: 104169. [PubMed: 31918042]

22. Guerra-Assunção JA, Houben RM, Crampin AC, et al. Recurrence due to relapse or reinfection with *Mycobacterium tuberculosis*: a whole-genome sequencing approach in a large, population-based cohort with a high HIV infection prevalence and active follow-up. J Infect Dis 2015; 211:1154–63. [PubMed: 25336729]

23. He W, Tan Y, Song Z, et al. Endogenous relapse and exogenous reinfection in recurrent pulmonary tuberculosis: a retrospective study revealed by whole genome sequencing. Front Microbiol 2023; 14:1115295. [PubMed: 36876077]

24. Korhonen V, Smit P, Haanperä M, et al. Whole genome analysis of Mycobacterium tuberculosis isolates from recurrent episodes of tuberculosis, Finland, 1995–2013. Clin Microbiol Infec 2016; 22:549–54. [PubMed: 27021423]

25. Li M, Qiu Y, Guo M, et al. Investigation on the cause of recurrent tuberculosis in a rural area in China using whole-genome sequencing: a retrospective cohort study. Tuberculosis 2022; 133:102174. [PubMed: 35124543]

26. Liu Q, Qiu B, Li G, et al. Tuberculosis reinfection and relapse in eastern China: a prospective study using whole-genome sequencing. Clin Microbiol Infec 2022; 28:1458–64. [PubMed: 35700940]

27. Mave V, Chen L, Ranganathan UD, et al. Whole genome sequencing assessing impact of diabetes mellitus on tuberculosis mutations and type of recurrence in India. Clin Infect Dis 2022; 75:768–76. [PubMed: 34984435]

28. Norrby M, Groenheit R, Mansjö M, et al. Whole genome sequencing of recurrent tuberculosis in Stockholm County 1996–2016. J Public Health Emerg 2020; 4:31.

29. Parvaresh L, Crighton T, Martinez E, Bustamante A, Chen S, Sintchenko V. Recurrence of tuberculosis in a low-incidence setting: a retrospective cross-sectional study augmented by whole genome sequencing. BMC Infect Dis 2018; 18:181–6. [PubMed: 29665796]

30. Sadovska D, Nodieva A, Pole I, et al. Advantages of analysing both pairwise SNV-distance and differing SNVs between *Mycobacterium tuberculosis* isolates for recurrent tuberculosis cause determination. Microb Genom 2023; 9:mgen000956. [PubMed: 36951900]

31. Shanmugam S, Bachmann NL, Martinez E, et al. Whole genome sequencing based differentiation between re-infection and relapse in Indian patients with tuberculosis recurrence, with and without HIV co-infection. Int J Infect Dis 2021; 113:S43–7. [PubMed: 33741489]

32. Witney AA, Bateson AL, Jindani A, et al. Use of whole-genome sequencing to distinguish relapse from reinfection in a completed tuberculosis clinical trial. BMC Med 2017; 15:151–13. [PubMed: 28793891]

33. Naidoo K, Dookie N. Insights into recurrent tuberculosis: relapse versus reinfection and related risk factors. Tuberculosis 2018. doi:10.5772/intechopen.73601

34. Lee RS, Radomski N, Proulx J-F, et al. Reemergence and amplification of tuberculosis in the Canadian Arctic. J Infect Dis 2015; 211:1905–14. [PubMed: 25576599]

35. Marx FM, Dunbar R, Enarson DA, et al. The temporal dynamics of relapse and reinfection tuberculosis after successful treatment: a retrospective cohort study. Clin Infect Dis 2014; 58:1676–83. [PubMed: 24647020]

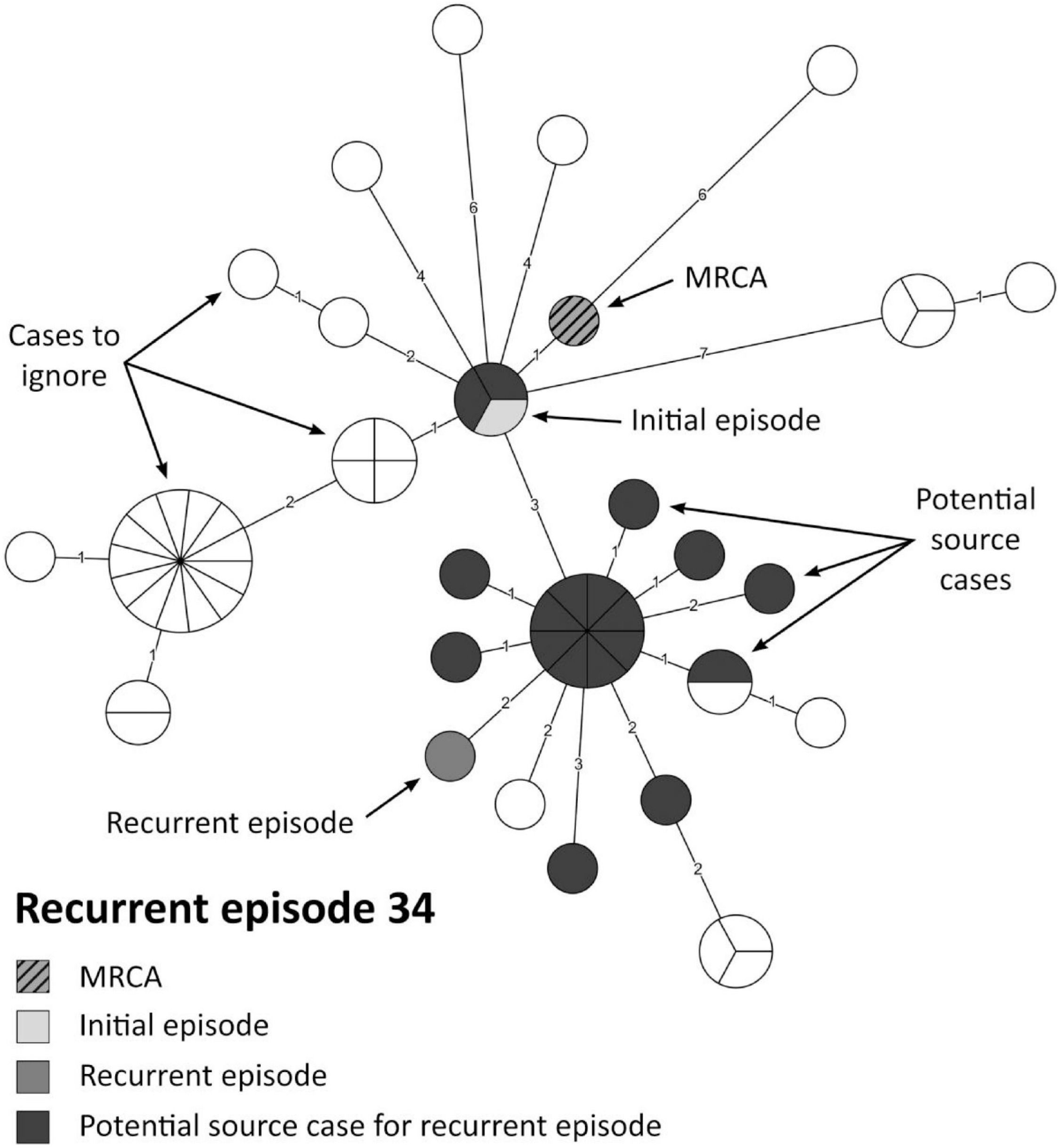

Figure 1.

Example of a phylogenetic tree created for a cluster containing 1 or more episodes of recurrent tuberculosis in American Indian or Alaska Native persons: Alaska, 2008–2020. *Mycobacterium tuberculosis* isolates associated with tuberculosis cases are represented by nodes (circles); ≥2 isolates with pairwise whole genome single-nucleotide polymorphism (wgSNP) distances of 0 are displayed together in a single segmented node. Numeric labels on branches denote wgSNP distances between nodes or segmented nodes. Nodes colored light gray and medium gray denote the initial and recurrent cases for the focal recurrent episode, respectively. Potential source cases (nodes colored dark gray) are cases that may have given rise to the recurrent case under a reinfection scenario; they include cases counted during the interepisode interval (period between the initial and recurrent case count dates, the latter extended 90 days to account for reporting delays) with an isolate having a pairwise wgSNP distance from the recurrent case isolate less than or equal to the wgSNP distance between the recurrent and corresponding initial case isolate (both distances based on the cluster wgSNP comparison). Nodes colored white correspond to isolates associated with cases that did not meet 1 or more criteria to be considered a potential source case and can be ignored. The node colored medium gray with black stripes denotes the most recent common ancestor (MRCA) for isolates in the tree. Phylogenetic trees were produced only for recurrent episodes for which the pairwise wgSNP distance between isolates associated with the recurrent and corresponding initial case was ≤10.

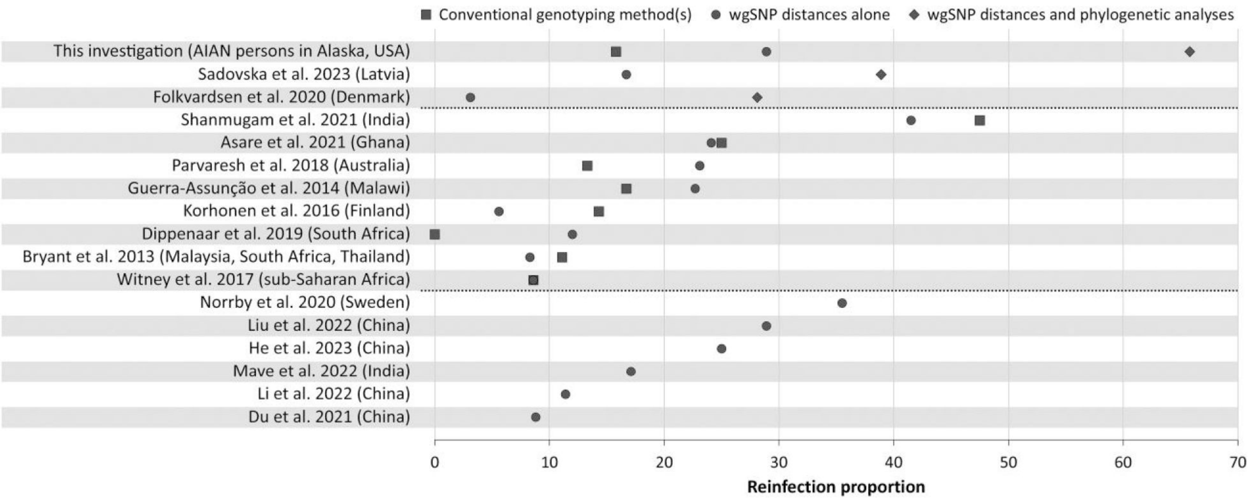

Figure 2.

Point estimates of reinfection proportions reported in, or enumerated based on findings of, 16 publications identified by a literature search for investigations that sought to distinguish recurrent tuberculosis relapse and reinfection etiologies using whole genome sequencing; results of this investigation are included for comparison. Investigations are identified on the left, with associated settings indicated parenthetically. Methods used to evaluate genetic distances between initial and recurrent episodes are indicated symbolically. Dashed lines divide the figure into 3 panels that present results of investigations that considered whole genome single-nucleotide polymorphism (wgSNP) distances and phylogenetic analyses and wgSNP distances alone (upper), both wgSNP distances alone and conventional genotyping methods (middle), and only wgSNP distances alone (lower). Within panels, investigations are arranged vertically in descending order by magnitude of largest reinfection proportion. For additional information about the associated investigations, see Supplementary Table 4; literature search conducted on 3 October 2023.

Table 1.

Criteria Used to Attribute the Etiology of 38 Episodes of rTB in AIAN Persons to Relapse or Reinfection Based on WGS Data: Alaska, 2008–2020

Criterion	Etiologic Attribution Method (WGS Data used)	Criterion Details	Etiologic Attribution	Rationale
1	wgSNP distances alone	Pairwise wgSNP distance[a] between isolates associated with the recurrent and corresponding initial case >10 [16].	Reinfection	Isolate genetic distance too large to be consistent with relapse.
1	wgSNP distances alone	Pairwise wgSNP distance[a] between isolates associated with the recurrent and corresponding initial case ≤10 [16].	Relapse	Isolate genetic distance small enough to be consistent with relapse.
2	wgSNP distances and phylogenetic analyses	Pairwise wgSNP distance[a] between isolates associated with the recurrent and corresponding initial case ≤10 AND at least 1 potential source case isolate more closely related to recurrent case isolate than initial case isolate (based on wgSNP distances) AND initial and recurrent case isolates separated by ≥2 branches in the phylogenetic tree.	Reinfection	Potential source of reinfection identified and is more likely than relapse based on tree.
3	wgSNP distances and phylogenetic analyses	Pairwise wgSNP distance[a] between isolates associated with the recurrent and corresponding initial case ≤10 AND no potential source case isolate more closely related to recurrent case isolate than initial case isolate (based on wgSNP distances).	Relapse	No potential source of reinfection identified.
4	wgSNP distances and phylogenetic analyses	Pairwise wgSNP distance[a] between isolates associated with the recurrent and corresponding initial case ≤10 AND at least 1 potential source case isolate as closely related to recurrent case isolate as initial case isolate (based on wgSNP distances).	Indeterminant	Potential source of reinfection identified but is equally as likely as relapse based on tree.
5	wgSNP distances and phylogenetic analyses	Pairwise wgSNP distance[a] between isolates associated with the recurrent and corresponding initial case ≤10 AND at least 1 potential source case isolate more closely related to recurrent case isolate as initial case isolate (based on wgSNP distances) AND initial and recurrent case isolates separated by 1 branch in the phylogenetic tree.	Indeterminant	Potential source of reinfection identified but some uncertainty of direction of transmission due to possibility of intrahost diversity in initial episode.
6	wgSNP distances and phylogenetic analyses	Pairwise wgSNP distance[a] between isolates associated with the recurrent and corresponding initial case ≤10 AND at least 1 potential source case isolate more closely related to recurrent case isolate as initial case isolate (based on wgSNP distances) AND the potential source cases counted concurrently (ie, in the same month) as the recurrent case.	Indeterminant	Potential source of reinfection too close in time to recurrent episode to establish direction of transmission.
7	wgSNP distances and phylogenetic analyses	Pairwise wgSNP distance[a] between isolates associated with the recurrent and corresponding initial case ≤10 AND no potential source case isolate more closely related to recurrent case isolate than initial case isolate (based on wgSNP distances) AND potential source case coverage associated with recurrent instance <60%.	Indeterminant	No potential source of reinfection identified but increased likelihood potential source could be missing from analysis.

Abbreviations: AIAN, American Indian or Alaska Native; rTB, recurrent tuberculosis; WGS, whole genome sequencing; wgSNP, whole genome single-nucleotide polymorphism.

[a] wgSNP distances from cluster wgSNP comparison associated with the cluster containing the recurrent case isolate.

Table 2.

TB Cases (Including Recurrent Cases) and TB Incidence Among US-Born Persons by Race and Ethnicity and Location Where Cases Were Counted: United States, 2008–2020

	TB Cases, No. (%)		Recurrent Cases, No. (%)		TB Cases Classified as Recurrent Cases, % (95% CI)[a]		TB Incidence (Cases per 100 000 Persons)[b]	
Race and Ethnicity Group[c]	**Alaska**	**US[d]**	**Alaska**	**US[d]**	**Alaska**	**US[d]**	**Alaska**	**US[d]**
AIAN	557 (91.8)	1647 (3.8)	66 (97.1)	116 (7.0)	11.8 (9.4–14.8)	7.0 (5.9–8.4)	31.3	2.5
NHPI	5 (0.8)	293 (0.7)	0	2 (0.1)	0	0.7 (0.2–2.5)	3.7	2.3
Hispanic	3 (0.5)	8145 (19.0)	0	200 (12.0)	0	2.5 (2.1–2.8)	0.4	1.8
Non-Hispanic								
Asian alone	6 (1.0)	1610 (3.8)	0	34 (2.0)	0	2.1 (1.5–2.9)	3.6	2.3
Black alone	6 (1.0)	16 974 (39.6)	0	768 (46.2)	0	4.5 (4.2–4.8)	1.6	3.5
White alone	28 (4.6)	13 990 (32.6)	1 (1.5)	538 (32.4)	3.6 (0.6–17.7)	3.8 (3.5–4.2)	0.5	0.5
Multiracial	0	139 (0.3)	-	3 (0.2)	-	2.2 (0.7–6.2)	0	0.2
Unknown	2 (0.3)	84 (0.2)	1 (1.5)	2 (0.1)	50.0 (9.5–90.6)	2.4 (0.7–8.3)	-	-
All cases	607	42882	68	1663	11.2 (8.9–14.0)	3.9 (3.7–4.1)	7.1	1.1

Abbreviations: AIAN, American Indian or Alaska Native; NHPI, Native Hawaiian or other Pacific Islander; TB, tuberculosis.

[a] 95% CIs calculated using the Wilson score interval method [13].

[b] TB incidence calculated during 2009 to 2020 only because 2008 population estimates were not available for all race and ethnicity groups. For each combination of race and ethnicity group and location where cases were counted, incidence during 2009 to 2020 was calculated as 100 000 times the quotient of the sum of TB cases across years and the sum of annual population estimates across years, as obtained from the US Census Bureau's American Community Survey 5-year public use microdata sample data set [14]. Population estimates were not available for persons with unknown race or ethnicity.

[c] Cases were assigned to 1 of 8 race and ethnicity groups per the following classification sequence applied to self-reported data: AIAN race (irrespective of ethnicity and including single-race and multiracial persons); NHPI race (irrespective of ethnicity and including single-race and multiracial persons); Hispanic ethnicity; Asian race alone, Black race alone, White race alone, or ≥2 races and non-Hispanic ethnicity; persons with unknown race or ethnicity.

[d] United States includes the 48 contiguous states, the District of Columbia, Hawaii, and Alaska.

Table 3.

Etiologic Attributions of 38 Episodes of rTB in AIAN Persons Based on wgSNP Distances and Phylogenetic Analyses, wgSNP Distances Alone, and Conventional Genotyping Methods: Alaska, 2008–2020

		wgSNP Distances Alone		Conventional Genotyping Methods	
wgSNP Distances and Phylogenetic Analyses		**Reinfection**	**Relapse**	**Reinfection**	**Relapse**
Reinfection	25	11	14	6	19
Relapse	2	0	2	0	2
Indeterminant[a]	11	0	11	0	11

Abbreviations: AIAN, American Indian or Alaska Native; rTB, recurrent tuberculosis; wgSNP, whole genome single-nucleotide polymorphism.

[a] The indeterminant category was applicable only for attributions based on wgSNP distances and phylogenetic analyses because of the methods chosen to define the distance-based thresholds for attributions based on wgSNP distances alone and on conventional genotyping methods.