Whole genome sequencing for tuberculosis in Victoria, Australia: A genomic implementation study from 2017 to 2020

Katie Dale,[a]* Maria Globan,[b] Kristy Horan,[c] Norelle Sherry,[c] Susan Ballard,[c] Ee Laine Tay,[d] Simone Bittmann,[a] Niamh Meagher,[e,f] David J. Price,[e,f] Benjamin P. Howden,[g,c] Deborah A. Williamson,[b,f,h] and Justin Denholm [a,g]

[a]Victorian Tuberculosis Program, Melbourne Health, at the Peter Doherty Institute for Infection and Immunity, Melbourne, Victoria, Australia
[b]Victorian Infectious Diseases Reference Laboratory (VIDRL), at the Peter Doherty Institute for Infection and Immunity, Melbourne, Victoria, Australia
[c]Microbiological Diagnostic Unit Public Health Laboratory, The University of Melbourne, at the Peter Doherty Institute for Infection and Immunity, Melbourne, Victoria, Australia
[d]Communicable Disease Epidemiology and Surveillance, Health Protection Branch, Public Health Division, Department of Health, Victoria, Australia
[e]Department of Infectious Diseases at the Doherty Institute for Infection & Immunity, The University of Melbourne and Royal Melbourne Hospital, Melbourne, Victoria, Australia
[f]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Victoria, Australia
[g]Department of Microbiology and Immunology, The University of Melbourne, at the Peter Doherty Institute for Infection and Immunity, Melbourne, Victoria, Australia
[h]Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia

Summary

Background Whole genome sequencing (WGS) is increasingly used by tuberculosis (TB) programs to monitor *Mycobacterium tuberculosis* (*Mtb*) transmission. We aimed to characterise the molecular epidemiology of TB and *Mtb* transmission in the low-incidence setting of Victoria, Australia, and assess the utility of WGS.

Methods WGS was performed on all first *Mtb* isolates from TB cases from 2017 to 2020. Potential clusters (≤12 single nucleotide polymorphisms [SNPs]) were investigated for epidemiological links. Transmission events in highly-related (≤5 SNPs) clusters were classified as likely or possible, based on the presence or absence of an epidemiological link, respectively. Case characteristics and transmission settings (as defined by case relationship) were summarised. Poisson regression was used to examine associations with secondary case number.

Findings Of 1844 TB cases, 1276 (69.2%) had sequenced isolates, with 182 (14.2%) in 54 highly-related clusters, 2–40 cases in size. Following investigation, 140 cases (11.0% of sequenced) were classified as resulting from likely/possible local-transmission, including 82 (6.4%) for which transmission was likely. Common identified transmission settings were social/religious (26.4%), household (22.9%) and family living in different households (7.1%), but many were uncertain (41.4%). While household transmission featured in many clusters (n = 24), clusters were generally smaller (median = 3 cases) than the fewer that included transmission in social/religious settings (n = 12, median = 7.5 cases). Sputum-smear-positivity was associated with higher secondary case numbers.

Interpretation WGS results suggest *Mtb* transmission commonly occurs outside the household in our low-incidence setting. Further work is required to optimise the use of WGS in public health management of TB.

Funding The Victorian Tuberculosis Program receives block funding for activities including case management and contact tracing from the Victorian Department of Health. No specific funding for this report was received by manuscript authors or the Victorian Tuberculosis Program, and the funders had no role in the study design, data collection, data analysis, interpretation or report writing.

The Lancet Regional Health - Western Pacific 2022;28: 100556
Published online xxx
https://doi.org/10.1016/j.lanwpc.2022.100556

*Corresponding author at: Victorian Tuberculosis Program, The Peter Doherty Institute of Infection and Immunity, Level 5, 792 Elizabeth St, Melbourne, VIC 3000, Australia.
E-mail addresses: katie.dale@mh.org.au, kwebby@hotmail.com (K. Dale).

Keywords: TB; public health; Molecular epidemiology; Transmission; Whole genome sequencing

Research in context

Evidence before this study

Whole genome sequencing (WGS) is increasingly used by tuberculosis (TB) programs to monitor *Mycobacterium tuberculosis* (*Mtb*) transmission, with its increased discriminatory power compared to previous typing methods allowing improved detection of clusters and chains of transmission. We searched Medline for all articles published from database inception to 31 December 2021, with the terms "*Mycobacterium tuberculosis*" and "whole genome sequencing", limited to human studies. Results revealed that WGS and epidemiological data have been used to describe clustering in many settings and to investigate *Mtb* transmission contexts in high-incidence settings. However, no studies have yet used WGS and epidemiological data to provide a detailed characterisation of *Mtb* transmission, including an examination of transmission by age-group and transmission setting, in a low-incidence setting. Additionally, no studies have assessed the utility of routine WGS to support TB program activities.

Added value of this study

Our study assesses the utility of WGS for TB program activities over four years (2017–2020) in the low-incidence setting of Victoria, Australia, and uses WGS and epidemiological data to characterise *Mtb* transmission. We found that *Mtb* transmission commonly occurs outside the household and, while this has been previously suggested by genotyping studies in high incidence settings, to our knowledge, ours is the first study to demonstrate this in a low-incidence setting. Transmission among young adults in social/religious settings accounted for the majority of the identified extra-household instances and the largest clusters. However, a significant proportion also occurred between individuals without identified epidemiological links, which limits the ability to translate WGS findings into public health interventions in our setting.

Implications of all the available evidence

Integrating WGS and epidemiological data can improve our understanding of *Mtb* transmission and so guide TB control and case-finding efforts. However, continued comprehensive investigations of instances of WGS clustering without epidemiological links will be important to advance our understanding.

Introduction

Tuberculosis (TB) is an infectious disease predominantly caused by the bacteria, *Mycobacterium tuberculosis* (*Mtb*).[1] Genotyping of *Mtb* isolates may be performed by TB reference laboratories to assist in monitoring disease transmission and outbreaks within populations. Genotyping is also used to identify laboratory cross-contamination events, distinguish recurrent disease episodes as relapses or reinfection, and classify *Mtb* isolates into lineages.[2–6] Molecular methods used for genotyping of *Mtb* have progressed from techniques such as spoligotyping,[7,8] restriction fragment length polymorphism (RFLP), variable number tandem repeat (VNTR)[9] and mycobacterial interspersed repetitive unit (MIRU) typing,[10] to the routine use of whole genome sequencing (WGS) in some centres.[11,12] WGS has increased discriminatory power compared to previous typing methods, allowing for the prediction of drug resistance[13] and improved detection of clusters and chains of transmission.[14]

The Australian state of Victoria has a population of approximately 6.6 million[15] and is considered to have a low TB incidence (7.1 per 100,000 in 2021[16,17]). A previous report assessing relatedness of *Mtb* in our setting (2003–2010) using MIRU-VNTR typing and epidemiological investigation concluded that 17.0% (390/2298) of cases 'possibly' resulted from local transmission (were genomically related), including 4.2% of cases that 'likely' did (were genomically related with known epidemiological links).[18] In Victoria, WGS has been routinely used to sequence all first isolates since 2018, and some prior isolates have also been retrospectively sequenced. Here, we present the genomic epidemiology of *Mtb* in Victoria from 2017 to 2020, characterising *Mtb* transmission in our low-incidence setting, and assessing the utility of genomic data integration into public health investigation.

Methods

Study population

The study population included all TB cases diagnosed in Victoria and notified to the Department of Health from 1 January 2017 to 31 December 2020, focusing on culture-confirmed cases with associated sequenced isolates. Study start and end dates were determined by the

commencement of WGS in our setting (2017) and data completeness, respectively. Notification of TB to public health authorities is mandatory for clinicians and laboratories in Australia. A confirmed case requires culture or polymerase chain reaction confirmation of *Mtb*, or clinical/radiological diagnosis by a medical practitioner experienced in TB management.[19] All notified cases of TB are followed up by the Victorian Tuberculosis Program (VTP), and epidemiological data are collected during contact tracing investigations.

Susceptibility testing, whole genome sequencing and bioinformatic analysis of Mtb complex isolates

During the study period, *Mtb* isolates were referred to the Mycobacterium Reference Laboratory (MRL) at the Victorian Infectious Diseases Reference Laboratory (VIDRL) for identification and phenotypic susceptibility testing (See Appendix 1 for methods). Sequencing and genomic analyses were performed at the Microbiological Diagnostic Unit Public Health Laboratory. Genomic DNA was extracted from *Mtb* isolates grown on solid culture utilising bead beating and ethanol precipitation as previously described.[20] Briefly, unique dual indexed libraries were prepared using the Nextera XT DNA sample preparation kit (Illumina). Libraries were sequenced on the Illumina NextSeq500/550 with 150-cycle paired end chemistry as described by the manufacturer's protocols. Sequences not meeting predefined quality control metrics (minimum average quality score 30, target sequencing depth ≥40X were resequenced).

Sequences were analysed using a custom bioinformatics pipeline for *Mtb*,[21] incorporating mycobacterial species identification, lineage calling, detection of antimicrobial resistance (AMR)-conferring single nucleotide polymorphisms (SNPs), phylogenetic analysis and identification of genomic clusters.

Mycobacterial species identification was performed by *k*-mer identification using the kraken2 tool[22] and minikraken.v2 database[23,24] to identify *Mtb* complex sequences, followed by use of the SNP-IT tool[25] to identify *Mtb* sequences; non-*Mtb* sequences were excluded from further analysis. Lineage calling and detection of AMR-conferring SNPs were performed using the tb-profiler tool, employing the ReSeqTB database.[26]

Phylogenetic analysis was performed by aligning all genomes to the H37Rv (v3) reference (a widely used strain in tuberculosis research[27]) and core genome SNP calling using snippy (v4.4.5),[28] phylogenetic tree building using iqtree (v1.6.12).[29] Pairwise SNP distances were also calculated from this alignment and single-linkage clustering performed on the resulting matrix. Genomic clusters were defined at two levels: ≤5 SNPs (highly related) and 6–12 SNPs (potentially related), as these thresholds are widely accepted/used internationally.[30–32]

Incorporation of genomic sequencing results into programmatic responses

Integration of whole genome sequencing results into clinical and public health practice involved monthly meetings and multilateral and continuous exchange between a multidisciplinary team, including laboratory scientists, bioinformaticians, epidemiologists, clinicians, and public health nurses.

The possibility a case was diagnosed due to laboratory contamination was considered for all clustered cases. The final decision was made by the medical director of the VTP, based on discussion with a multidisciplinary team described above and consideration of all available evidence including: laboratory handling records, to identify opportunity for co-handling or co-process in specimens or isolates; medical opinion regarding the likelihood the patient had TB; and the presence or absence of epidemiological links between the cases.

Epidemiological classification of cases

Classification was conducted according to national cluster and outbreak definitions,[33] with additional definitions and methodological detail available in Appendix 2 and Box 1. Briefly, local-transmission (defined as transmission occurring within the state of Victoria) was classified as 'likely' where both epidemiological and genomic links were demonstrated between two cases, 'possible' where genomic clustering was seen without a known epidemiological link, and 'probable' where epidemiological links were identified but genomic data were unavailable. In each instance, transmission between the two cases also needed to be physically, geographically and temporally possible (Appendix 2 and Box 1). An epidemiological link was defined as when exposure to a suspected source case/s in a shared space was either known, or likely based on identified shared personal connections, when the suspected source case/s were infectious. Characteristics of suspected instances of transmission were analysed, including case demographics and transmission settings. When describing the age characteristics of source cases, for instances where a secondary case had more than one possible source case, the average age of all possible source cases was presented. Findings in highly related (≤5 SNPs) and potentially related (6–12 SNPs) clusters are separately presented. Sequenced isolates from cases notified prior to the study period were considered when classifying cases and when referring to cluster size, but those occurring after were not.

Box 1 Definitions of commonly used terms in this report

Case: Any person with tuberculosis (TB) who is diagnosed and notified to the Victorian Department of

Health. A person may be notified more than once due to relapse or reinfection, and so each person may result in more than one case. In this manuscript, all of the 182 cases in highly related clusters during the study period were separate individuals.

Isolate: A pure culture of *Mycobacterium tuberculosis* grown from a pulmonary or extra-pulmonary sample taken from a TB case. The first isolate for each TB case was used in this study.

Clusters:

Highly related: two or more cases with *Mtb* isolates that differ by ≤5 SNPs on genomic analysis.

Potentially related: two or more cases with *Mtb* isolates that differ by 6-12 SNPs on genomic analysis.

Instances of local transmission (see Appendix 2 for a more detailed explanation):

Likely: suspected secondary and source case/s share an epidemiological link (see below) and also have *Mtb* isolates that are highly genomically-related (differ by ≤5 SNPs) on WGS.

Likely – source assigned: the "likely source case" for a suspected secondary case can be assigned because the suspected secondary case only had a known epidemiological link with one other pulmonary case in the cluster.

Likely – uncertain source: the source case is uncertain because more than one possible source case has genomic and epidemiological links with a secondary case.

Possible: *Mtb* isolates of suspected secondary and source case/s are highly genomically-related (differ by ≤5 SNPs) on WGS and no epidemiological link (see below) could be identified between the two cases. Additionally, transmission was physically, temporally and geographically possible, i.e. the possible source case was pulmonary and over the age of nine (i.e. potentially infectious with adult type pulmonary disease[34,35]); preceded the possible secondary case, with reference to the potential infectious period of the earlier source case/s and the symptom onset of the secondary case; and both cases were present in our setting when the source case was potentially infectious.

Probable: suspected secondary case has a known epidemiological link (see below) with a suspected source case, but whole genome sequencing data was unavailable for one or both of the cases, because, for example, the secondary case was culture negative.

Impossible: where transmission could not have occurred between two cases because the first case was extra pulmonary, or because the two cases had no relevant temporal overlap in Australia, e.g. the suspected secondary case migrated to Australia after the suspected source case completed treatment.

Epidemiological link: a suspected secondary case is known to have had contact with a suspected source case/s in a shared space, or this is likely based on identified shared personal connections. Temporally, this contact occurred when the suspected source case/s were potentially infectious, and prior to the symptom onset of the secondary case. Exposures of any duration were considered.

Transmission settings (defined by the relationship between the likely/possible source and secondary cases)**:**

Household: likely source and secondary cases share the same address.

Family (living in different households): likely source and secondary cases are members of the same family but don't share the same address.

Social: likely source and secondary cases are friends living at separate addresses.

Religious: likely source and secondary cases attended the same religious setting and live at separate addresses.

Unknown: the transmission setting was uncertain/unknown, due to the absence of known epidemiological links.

Extra-household: All categories other than 'household' are regarded as extra-household transmission.

Abbreviations: Mtb = Mycobacterium tuberculosis; SNP = single nucleotide polymorphisms; WGS = whole genome sequencing.

Statistical analyses, including associations between TB case characteristics and Mtb transmission

Differences between the characteristics of cases that were sequenced and unsequenced were assessed using Chi-Squared tests. The agreement between phenotypic and genotypic antimicrobial resistance determinations was assessed using Cohen's Kappa coefficient, and we report 95% confidence intervals.

Poisson regression, was used to examine associations between case characteristics and the number of likely secondary cases, with an offset representing time in the study and a random effect term used to model variation between clusters (and, indirectly, lineage). Multivariable logistic regression was also used to examine associations between case characteristics and two outcomes of interest: being a likely and/or possible source case (including cluster size as an independent variable and cluster as a random effect term); and having pulmonary involvement. All analyses were performed using R, version 4.2.0.[36] Cases missing one or more characteristics were excluded from the analysis. Univariable regression analyses were performed and reported to show the unadjusted relationship between the independent and dependent variables. The multivariable model was built using a backwards stepwise procedure based on Akaike information criterion. No adjustments were made for multiple testing.

Role of funding source

The VTP receives block funding for activities from the Victorian Department of Health. No specific funding for this report was received by manuscript authors or

the VTP, and the funders had no role in the study design, data analysis, interpretation or report writing.

Results

A total of 1854 TB cases were notified in Victoria during the study period. Culture-confirmed isolates were obtained from 1392 (1392/1854; 75.1%), and 1286 (1286/1854; 69.4%, 92.4% of culture confirmed) underwent WGS (Figure 1 and Table 1). Following WGS, ten cases were de-notified as they were determined to be due to laboratory contamination, reducing the number of notified cases to 1844 and the number of sequenced cases to 1276 (69.2%). The proportion of sequenced isolates varied across the study years (70.5%, 98.5%, 99.4%

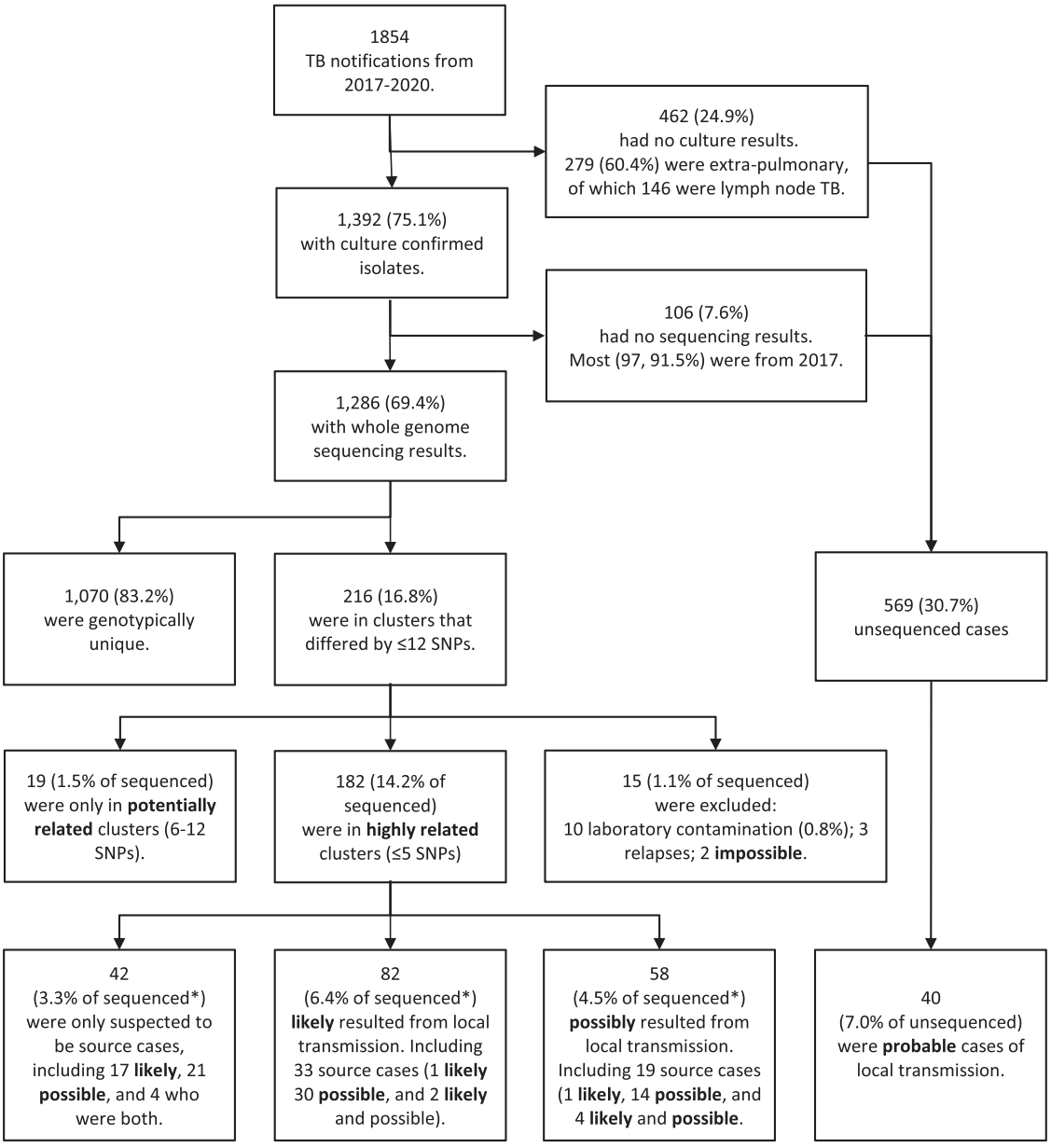

Figure 1. Flow chart showing the results of molecular and epidemiological investigations into TB cases from 2017 to 2020 in Victoria. See Box 1 of the main text for definitions. *The denominator for these calculations is 1,276, and excludes nine notified cases determined to be instances of laboratory contamination.

	Cases with sequenced isolates	Clustered (≤5 SNPs)	Cluster sizes (range)	Cases resulting from local transmission		Source cases	
				Likely	Likely and possible	Likely	Likely and possible
Total	1276	206	(2 −40)	82	140	29	94
Year							
2017	232 (18.2%)	39 (18.9%)	(2−31)	10 (12.2%)	22 (15.7%)	9 (31%)	27 (28.7%)
2018	337 (26.4%)	68 (33%)	(2 −33)	30 (36.6%)	44 (31.4%)	11 (37.9%)	25 (26.6%)
2019	341 (26.7%)	55 (26.7%)	(2−36)	29 (35.4%)	39 (27.9%)	6 (20.7%)	24 (25.5%)
2020	366 (28.7%)	44 (21.4%)	(2 −40)	13 (15.9%)	35 (25%)	3 (10.3%)	18 (19.1%)
Sex							
Female	707 (55.4%)	110 (53.4%)	(2 −40)	45 (54.9%)	76 (54.3%)	14 (48.3%)	56 (59.6%)
Male	569 (44.6%)	96 (46.6%)	(2−39)	37 (45.1%)	64 (45.7%)	15 (51.7%)	38 (40.4%)
Age group							
0-4	15 (1.2%)	13 (6.3%)	(2−20)	8 (9.8%)	13 (9.3%)	0.00	1 (1.1%)
5-14	21 (1.6%)	12 (5.8%)	(2 −21)	9 (11%)	12 (8.6%)	0.00	4 (4.3%)
15-24	252 (19.7%)	56 (27.2%)	(2 −39)	33 (40.2%)	44 (31.4%)	8 (27.6%)	32 (34%)
25-34	374 (29.3%)	52 (25.2%)	(2 −40)	17 (20.7%)	33 (23.6%)	7 (24.1%)	21 (22.3%)
35-64	400 (31.3%)	54 (26.2%)	(2−37)	9 (11%)	26 (18.6%)	11 (37.9%)	29 (30.9%)
65+	214 (16.8%)	19 (9.2%)	(2−38)	6 (7.3%)	12 (8.6%)	3 (10.3%)	7 (7.4%)
Median age	34 (25 - 55)	29 (22 - 43)	NA	23.5 (18 - 31.75)	25 (19 - 37)	33 (23 - 51)	29 (22.25 - 43)
Manifestation							
Extra Pulmonary	394 (30.9%)	33 (16%)	(2−35)	11 (13.4%)	25 (17.9%)	0	0
Pulmonary	628 (49.2%)	130 (63.1%)	(2−38)	58 (70.7%)	84 (60%)	23 (79.3%)	70 (74.5%)
Pulmonary Plus Other Sites	254 (19.9%)	43 (20.9%)	(2−40)	13 (15.9%)	31 (22.1%)	6 (20.7%)	24 (25.5%)
Sputum Smear							
Negative/ Unknown	1007 (78.9%)	137 (66.5%)	(2−38)	60 (73.2%)	106 (75.7%)	7 (24.1%)	43 (45.7%)
Positive	269 (21.1%)	69 (33.5%)	(2−40)	22 (26.8%)	34 (24.3%)	22 (75.9%)	51 (54.3%)
Cavity							
No results	318 (24.9%)	35 (17%)	(2−33)	12 (14.6%)	27 (19.3%)	5 (17.2%)	8 (8.5%)
No	761 (59.6%)	118 (57.3%)	(2−40)	53 (64.6%)	86 (61.4%)	10 (34.5%)	48 (51.1%)
Yes	197 (15.4%)	53 (25.7%)	(2−39)	17 (20.7%)	27 (19.3%)	14 (48.3%)	38 (40.4%)
Cough							
Missing	16 (1.3%)	2 (1%)	(2−5)	2 (2.4%)	2 (1.4%)	0	0
No	694 (54.4%)	98 (47.6%)	(2−40)	44 (53.7%)	74 (52.9%)	5 (17.2%)	32 (34%)
Yes	566 (44.4%)	106 (51.5%)	(2 −38)	36 (43.9%)	64 (45.7%)	24 (82.8%)	62 (66%)
Symptoms							
Missing	16 (1.3%)	2 (1%)	(2− 5)	2 (2.4%)	2 (1.4%)	0	0
No	123 (9.6%)	16 (7.8%)	(2−21)	8 (9.8%)	10 (7.1%)	1 (3.4%)[b]	6 (6.4%)
Yes	1137 (89.1%)	188 (91.3%)	(2−40)	72 (87.8%)	128 (91.4%)	28 (96.6%)	88 (93.6%)
Resistance							
Missing	2 (0.2%)	0	0	0	0	0	0
Fully sensitive	1150 (90.1%)	190 (92.2%)	(2 −40)	76 (92.7%)	129 (92.1%)	27 (93.1%)	89 (94.7%)
Yes	124 (9.7%)	16 (7.8%)	(2−5)	6 (7.3%)	11 (7.9%)	2 (6.9%)	5 (5.3%)
Work							
Missing	48 (3.8%)	4 (1.9%)	(2 −23)	1 (1.2%)	3 (2.1%)	0.00	1 (1.1%)
Employed	524 (41.1%)	80 (38.8%)	(2−39)	29 (35.4%)	50 (35.7%)	11 (37.9%)	40 (42.6%)
Home duties	93 (7.3%)	26 (12.6%)	(2−22)	13 (15.9%)	23 (16.4%)	2 (6.9%)	5 (5.3%)
Retired	189 (14.8%)	17 (8.3%)	(2−7)	6 (7.3%)	9 (6.4%)	5 (17.2%)	8 (8.5%)
Student	279 (21.9%)	42 (20.4%)	(2−27)	19 (23.2%)	31 (22.1%)	3 (10.3%)	18 (19.1%)
Tourist/visitor	14 (1.1%)	1 (0.5%)	(2−1)	0	0	1 (3.4%)	1 (1.1%)
Unemployed	129 (10.1%)	36 (17.5%)	(2 − 40)	14 (17.1%)	24 (17.1%)	7 (24.1%)	21 (22.3%)

Table 1 *(Continued)*

	Cases with sequenced isolates	Clustered (≤5 SNPs)	Cluster sizes (range)	Cases resulting from local transmission		Source cases	
				Likely	Likely and possible	Likely	Likely and possible
Risk factors							
Substance abuse	35 (2.7%)	20 (9.7%)	(2– 40)	12 (14.6%)	18 (12.9%)	4 (13.8%)	15 (16%)
Ever homeless	31 (2.4%)	8 (3.9%)	(2–37)	2 (2.4%)	4 (2.9%)	2 (6.9%)	6 (6.4%)
Ever resided in prison	32 (2.5%)	10 (4.9%)	(2–38)	3 (3.7%)	5 (3.6%)	2 (6.9%)	8 (8.5%)
Australian-born child <15years, parent/s from high risk country	17 (1.3%)	9 (4.4%)	(2–21)	8 (9.8%)	9 (6.4%)	0.00	1 (1.1%)
Household member or close contact	203 (15.9%)	90 (43.7%)	(2–39)	66 (80.5%)	78 (55.7%)	7 (24.1%)	31 (33%)
Residency							
Unknown	19 (1.5%)	0	0	0	0	0	0
Australia born	107 (8.4%)	51 (24.8%)	(2–39)	28 (34.1%)	44 (31.4%)	3 (10.3%)	20 (21.3%)
Permanent resident	699 (54.8%)	124 (60.2%)	(2–40)	46 (56.1%)	82 (58.6%)	22 (75.9%)	59 (62.8%)
Overseas student	228 (17.9%)	18 (8.7%)	(2– 5)	3 (3.7%)	7 (5%)	3 (10.3%)	10 (10.6%)
Refugee / humanitarian	24 (1.9%)	5 (2.4%)	(2– 9)	3 (3.7%)	4 (2.9%)	0	2 (2.1%)
Unauthorized person	4 (0.3%)	0	0	0	0	0	0
Visitor	79 (6.2%)	3 (1.5%)	(2 – 29)	1 (1.2%)	2 (1.4%)	1 (3.4%)	1 (1.1%)
Other	116 (9.1%)	5 (2.4%)	(2– 8)	1 (1.2%)	1 (0.7%)	0	2 (2.1%)
Place of birth							
Missing	3 (0.2%)	0	0	0	0	0	0
Australian born	108 (8.5%)	51 (24.8%)	(2–39)	28 (34.1%)	44 (31.4%)	3 (10.3%)	20 (21.3%)
Overseas born	1165 (91.3%)	155 (75.2%)	(2 –40)	54 (65.9%)	96 (68.6%)	26 (89.7%)	74 (78.7%)
Median time to healthcare presentation (IQR)	27 (3–67)	30 (8 - 73)	NA	31 (11.5 - 70)	27 (3–75.5)	30 (12.5–61.75)	29 (12–73)
Median years since arrival before TB event date, if overseas-born (IQR)	5 (2 –14)	9 (3–16)	NA	13.5 (6.25 –16.75)	11.5 (4.25–16)	7 (3–14.75)	10 (3 –16)

***Table 1*: Characteristics of TB cases with sequenced isolates in Victoria 2017–2020[a].**

Abbreviations: SNP=single nucleotide polymorphisms; MDR=multi-drug resistant; IQR=interquartile range; NA=not applicable.

[a] Excludes 10 initially notified TB cases that were subsequently determined to have resulted from laboratory contamination.

[b] Although "No" was entered in the symptom field, the case was noted to have had a cough on-and-off for several months in their case notes.

and 99.5% in 2017, 2018, 2019 and 2020 respectively).

The 569 cases without sequenced isolates were made up of 462 culture-negative cases, and 107 culture-positive cases with unsequenced isolates, 97 (97/107; 90.7%) of which were from the end of 2017, when funding for sequencing was limited. Overall, in the unsequenced group there was a higher proportion of children under five years of age (24/569 [4.2%] versus 15/1276 [1.2%], p <0.001) and 5–14 years of age (18/569 [3.2%] versus 21/1276 [1.6%], p = 0.069) and a higher proportion of extrapulmonary cases (315/569 [55.3%] versus 395/1276 [31.0%], p < 0.001) compared to sequenced cases (Appendix Table 1).

Strain diversity

The 1276 *Mtb* isolates included in the study were from five different previously described lineages,[37] with Lineage 1 (Indo-Oceanic, 372/1276; 29·2%), Lineage 2 (East-

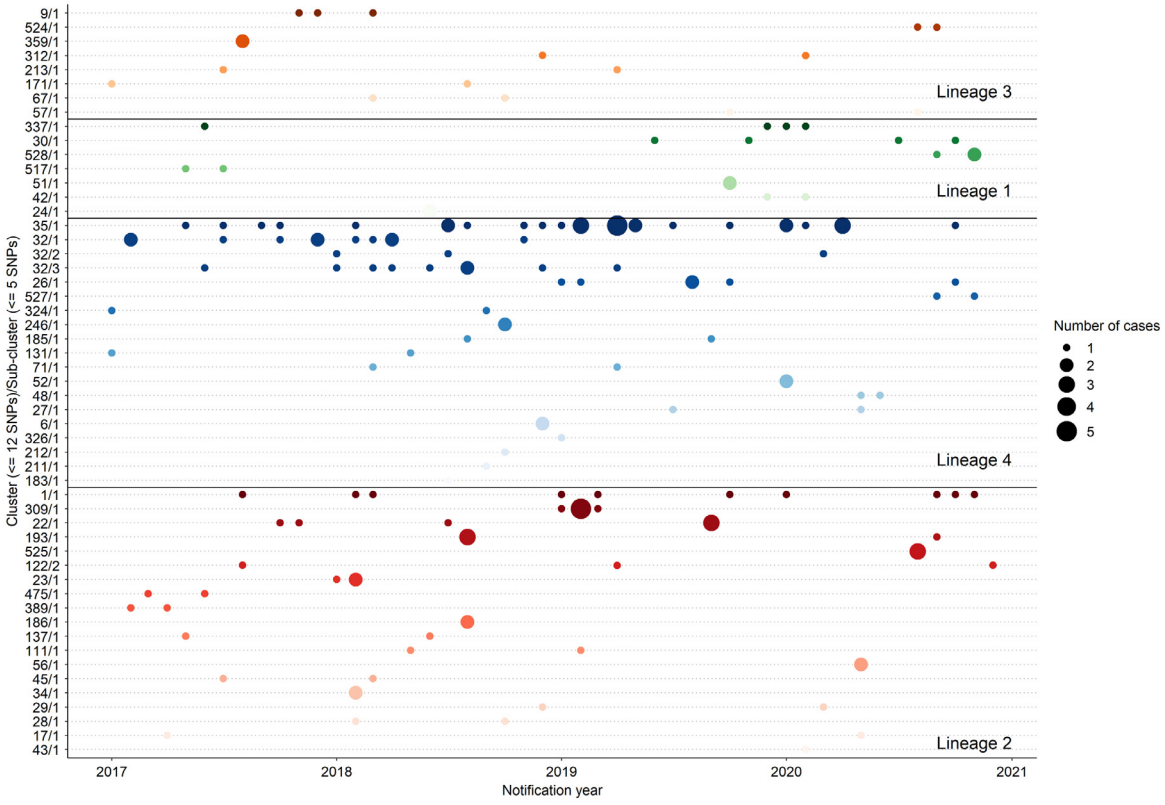

Figure 2. Timeline of genotypic clusters seen among Victorian TB cases from 2017 to 2020. Clusters are ordered on the *y* axis by lineage, cluster (≤ 12) size, and sub-cluster number (≤ 5), and this is also indicated by the lightening point colours. Lineages are given by the point colours: Lineage 3 = orange; Lineage 1 = green; Lineage 4 = blue, and Lineage 2 = red. On the y axis, the figures prior to and after the forward-slash refer to the cluster (≤ 12) and sub-cluster (≤ 5) numbers, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Asian, 333/1276; 26.1%), Lineage 3 (East-African-Indian, 236/1276; 18.5%) and Lineage 4 (Euro-American, 331/1276; 25.9%) accounting for 99.8% of all identified strains (Figure 2). There was also a single isolate from Lineage 6, a single *Mycobacterium bovis* isolate, and lineage could not be assigned for a further two isolates.

Phenotypic-genotypic concordance of antimicrobial resistance profiles

Phenotypic drug susceptibility testing results were obtained from 1383 isolates during the study period (excluding isolates determined to have resulted from laboratory contamination), with 112 (112/1383; 8.1%) found to be resistant to isoniazid, 23 (23/1383; 1.7%) resistant to rifampicin, 11 (11/1383; 0.8%) resistant to ethambutol, 28 (28/1383; 2.0%) resistant to pyrazinamide, and 11 (11/1383; 0.8%) resistant to moxifloxacin. There were eight (8/1383; 0.6%) multi-drug resistant cases (resistant to isoniazid and rifampicin) and nine (9/1383; 0.7%) pre-extensively drug-resistant cases (resistance to isoniazid, rifampicin and moxifloxacin or ofloxacin) during the study period.

Phenotypic and genotypic results for antimicrobial resistance were predominantly concordant among sequenced isolates. Where results were available, concordance was highest for rifampicin (1231/1233; 99·8%) and ethambutol (1226/1233; 99.4%), and lower for pyrazinamide (1216/1235; 98.5%) and isoniazid (1216/1234; 98.5%) (Table 2). Phenotypic and/or genotypic results were missing for the four main drugs for 9.7% of sequenced isolates.

Leading concordant mutations were katG (p.Ser315Thr) for isoniazid (61/1234; 4.9%), rpoB (p.Ser450Leu) for rifampicin (7/1233; 0.6%), pncA (p.Ser104Arg) for pyrazinamide (2/1235; 0.2%), and embB (p.Gln497Arg) for ethambutol (2/1233; 0.2%) (Appendix Table 2). The most common discordant gene mutations were fabG1 (c.-15C>T) for isoniazid (7/1234; 0.6%), a promoter mutation that is known to result in borderline resistance to isoniazid,[38] and rpoB (p.Leu452Pro) for rifampicin (1/1233; 0.1%). (Appendix Table 2).

Antibiotic	Number of isolates with genotyping and phenotyping results	Number with concordant results (proportion)	Number with discordant results (proportion)	Cohen's kappa statistic (95% confidence intervals)	Concordant (proportion)		Discordant(proportion)	
					Susceptible	Resistant	Phenotypically sensitive, Genotyping predicted resistance	Phenotypic resistance, genotyping predicted no resistance
Isoniazid	1234	1216 (98.5%)	18 (1.5%)	0.9 (0.9–0.9)	1121 (90.8%)	95 (7.7%)	14 (1.1%)	4 (0.3%)
Rifampicin	1233	1231 (99.8%)	2 (0.2%)	0.9 (0.9–1)	1213 (98.4%)	18 (1.5%)	1 (0.1%)	1 (0.1%)
Pyrazinamide	1235	1216 (98.5%)	19 (1.5%)	0.6 (0.3–0.8)	1204 (97.5%)	12 (1%)	7 (0.6%)	12 (1%)
Ethambutol	1233	1226 (99.4%)	7 (0.6%)	0.7 (0.5–0.9)	1217 (98.7%)	9 (0.7%)	7 (0.6%)	0
Moxifloxacin	132	130 (98.5%)	2 (1.5%)	0.9 (0.7–1)	123 (93.2%)	7 (5.3%)	0	2 (1.5%)

Table 2: **Concordance and discordance of genotyping and phenotyping results for antimicrobial resistance from Victorian TB cases with sequenced isolates, 2017–2020 (n = 1276, excludes instances of laboratory contamination). Results for all other antibiotics and specific gene mutations are shown in Appendix Table 4.**

Clustering

Of all 1286 original TB notifications with sequenced isolates during the study period, 216 (216/1286; 16.8%) were in 66 potentially related clusters (≤12 SNP differences). Each genomic cluster could be divided into one or more highly-related clusters (≤5 SNP difference), creating 80 highly related clusters (≤5 SNP difference) in total. There were 1070 (1070/1286; 83.2%) genotypically unique isolates.

After epidemiologic investigation, 15 cases within the highly-related clusters (≤5 SNP) were excluded as they were identified as relapse (n = 3) or local transmission was determined to be 'impossible' (n = 2) and, as mentioned earlier, ten cases were determined to be due to laboratory contamination (see Box 1 and Figure 1). A further 19 cases were only in potentially related clusters (6–12 SNPs). This left 182 cases (182/1276; 14.2% of sequenced cases) in 54 highly related clusters (≤5 SNP), ranging from two to 40 cases in size (the largest cluster includes 30 cases notified prior to the study period) (Figure 2). A sample of a genomic report used programmatically during the study period is included in Appendix Figure 1.

Of these 182 cases in highly related clusters (≤5 SNP) during the study period, 42 were determined to be exclusively likely and/or possible source cases (rather than from local transmission themselves) and 140 were secondary cases. Additionally, 52 of the 140 secondary cases were also likely and/or possible source cases themselves (i.e. they were potentially infectious and possibly/likely transmitted onwards). Of all 94 source cases, 29 were determined to be likely, as they were involved in likely instances of transmission (with known epidemiological links), and a further 65 were determined to be possible source cases, because they occurred prior to other cases in their cluster and, although they didn't share an epidemiological link with any subsequent cases, transmission was physically, temporally and geographically possible (see definitions in Box 1 and Appendix 2). Of the 140 secondary cases (140/1276; 11.0% of sequenced cases, ranging from 9.5–13.1% of annual cases), 82 cases (82/1276; 6.4% of sequenced cases) were designated as likely resulting from local transmission (with known epidemiological links, see definitions in Box 1) and 58 (58/1276; 4.5% of sequenced cases) possibly resulted from local transmission (without known epidemiological links). Additionally, there were 40 probable cases of transmission during the study period, identified based only on epidemiological links, because genomic results were unavailable for one or both cases.

Characteristics of local transmission

Considering all likely and possible instances of local transmission identified in highly-related clusters, 26.4% (37/140) occurred in social or religious settings, 22.9% (32/140) within households and 7.1% (10/140)

	Univariate analysis				Multivariate analysis			
	Incidence rate ratio	95% confidence interval	p value	Likelihood ratio test	Incidence rate ratio	95% confidence interval	*p* value	Likelihood ratio test
Event year (versus 2017)								
2018	1.62	(0.78−3.37)	0.195	0.082				
2019	2.59	(1.25−5.38)	0.011					
2020	1.61	(0.52−4.99)	0.410					
Sex (males versus females)	0.72	(0.42−1.24)	0.232	0.231	0.41	(0.18−0.92)	0.031	0.030
Age group, years (versus 35−65 years)								
0−14[c,b]								
15−24	1.01	(0.50−2.02)	0.987					
25−34	0.77	(0.39−1.50)	0.442					
65+	0.37	(0.13−1.08)	0.069					
Overseas-born (versus Australian born)	0.71	(0.30−1.66)	0.427	0.448	4.46	(1.00−19.85)	0.050	0.023
Work status (versus employed)								
Home duties	0.52	(0.12−2.25)	0.384	<0.001				
Retired	0.72	(0.27−1.92)	0.509					
Student	0.30	(0.09−1.02)	0.054					
Tourist/Visitor	5.82	(1.35−24.97)	0.018					
Unemployed	4.24	(2.28−7.88)	<0.001					
Residency status (versus Australian born)								
Permanent resident	0.99	(0.42−2.34)	0.985	<0.001				
Overseas student	0.25	(0.06−0.99)	0.048					
Refugee/humanitarian[b]								
Visitor	0.62	(0.12−3.07)	0.557					
Other[b]								
Ever homeless	8.85	(4.17−18.80)	<0.001	<0.001				
History of substance abuse	6.21	(3.03−12.74)	<0.001	<0.001				
Ever in a correctional facility	5.82	(2.49−13.63)	<0.001	0.001				
CXR suggestive of past TB	3.25	(1.74−6.09)	<0.001	<0.001	3.41	(0.99−11.80)	0.052	0.068
Lineage (versus Lineage 1)				<0.001				
Lineage 2	11.81	(2.79−49.98)	<0.001	<0.001				
Lineage 3	2.29	(0.38−13.70)	0.364					
Lineage 4	11.67	(2.75−49.49)	<0.001					
Lineage 6[b]								
Symptoms	6.00	(0.83−43.43)	0.076	0.015				
Cough	5.22	(2.55−10.72)	<0.001	<0.001	5.03	(1.60−15.83)	0.006	0.001
Cavity	5.23	(3.03−9.01)	<0.001	<0.001				
Sputum Smear positive	20.52	(9.66−43.58)	<0.001	<0.001	12.74	(4.44−36.60)	<0.001	<0.001
Any resistance to anti-microbials	0.97	(0.35−2.70)	0.957	0.957				

***Table 3*: Factors associated with the number of secondary cases that a TB case leads to in Victoria 2017−2020, among all sequenced cases. Only secondary cases arising from likely source cases were considered in this analysis.**[a]

[a] All analyses were performed with an offset for time in the study. Some TB cases ($n = 64$, 5.0%) were removed from the analysis due to missing data. The number of cases missing certain fields (not mutually exclusive) were, as follows: work ($n = 48$), lineage ($n = 3$), symptoms ($n = 16$), cough ($n = 16$), sputum smear result ($n = 5$), resistance information ($n = 2$). We did not account for this missing data.

[b] Due to no observed transmission events from cases in this group the estimated incidence rate ratio, 95% confidence interval and *p*-value are not meaningful, and therefore not reported.

[c] 0-14 year olds ($n = 34$) were removed from the multivariate model because there were no outcomes.

Note: We excluded the risk factor "Australian-born child <15years, parent/s from high risk country" as it perfectly predicted not being a source case. We also excluded "Household member or close contact" because it reflects such a variety of exposures, recent or not, household or not. We double-checked the impact of this variable in the multivariable model, and its impact was limited.

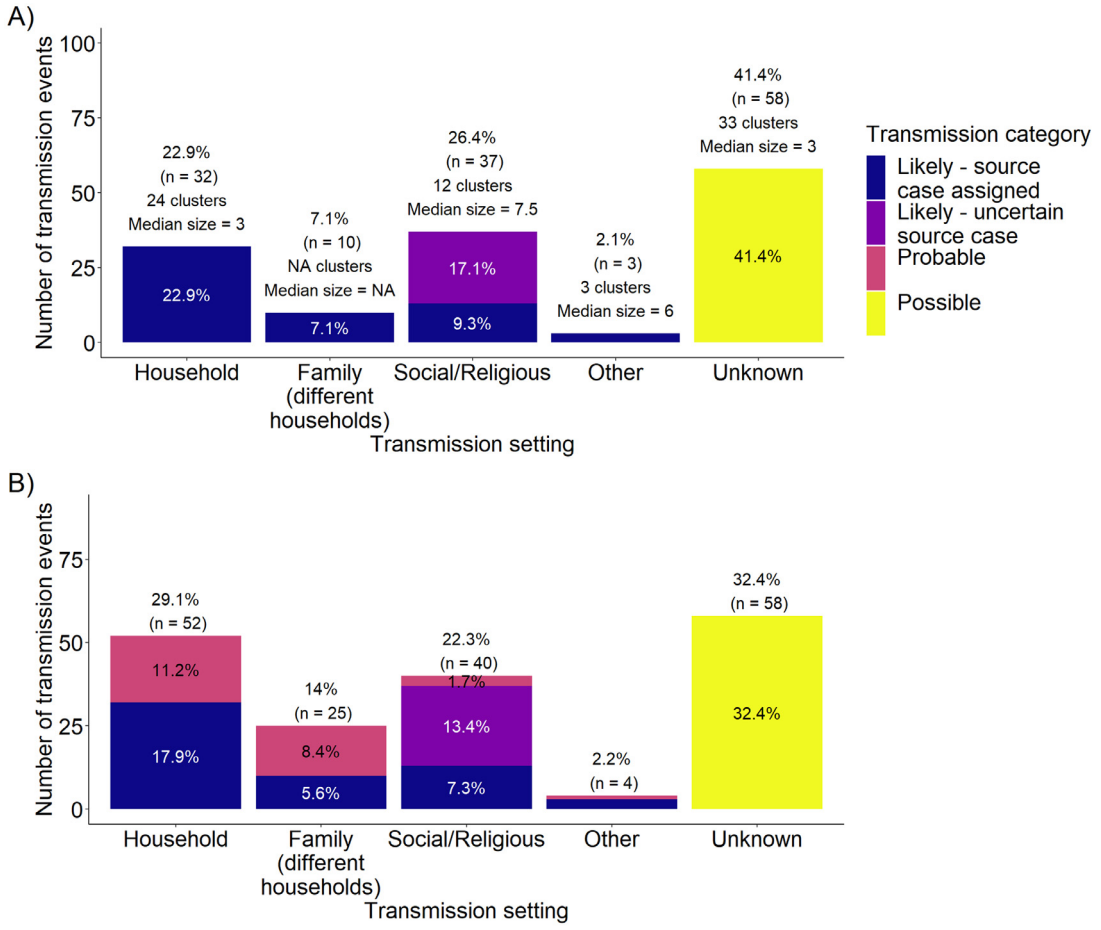

Figure 3. Proportion of transmission that occurred in different contexts for the 140 identified instances in Victoria 2017-2020, including instances that were A) likely or possible, or B) likely, possible or probable. "Other" includes instances of transmission in health, education and child care settings.

within families living in different households, henceforth simply referred to as "family" (Figure 3A). While household transmission featured in more clusters (n = 24) than transmission in social/religious settings (n = 12), the clusters were generally smaller (median of 3 cases, compared to 7.5 cases in social/religious settings). There were also two instances of likely transmission in a healthcare setting, and one in an education setting ("other" settings, 3/140; 2.1%). The remaining transmission settings were unknown (58/140, 41.4%). Therefore, if all 140 instances represent local transmission, 77.1% (108/140) occurred beyond household contacts (70.0% beyond household and family contacts; 98/140), and 82.4% (28/34) of all sputum smear positive secondary cases (i.e. those likely to be infectious themselves) resulted from transmission outside the household. The contribution of household transmission increased if probable instances (n = 40) were added (Figure 3b), with 70.6% (127/180) occurring beyond the household and 56.7% (102/180) beyond the household and family contacts. Transmission in household (21/40, 52.5%) and family (15/40, 37.5%) settings was common among the probable cases of transmission, with children under five and 5–14 years of age making up 40.0% (16/40) and 22.5% (9/40) of secondary cases, respectively. Overseas born children <15 years of age made up 10.0% (4/40) of probable secondary cases, and 47.5% (19/40) were Australian-born children with parent/s from a high-risk country.

The 58 possible cases of local transmission (with unknown epidemiological links) accounted for all instances where the transmission setting was unknown. Of these, all shared one or more of the following characteristics with possible source case/s in their clusters, including: social risk factors (e.g. substance use, 13/58, 22.4%), religious setting attendance (10/58, 17.2%), local government area residence (25/58, 43.1%), diagnosis within the same year (n = 39/58, 67.2%), or country of birth (11/58,

19.0%); and 22 cases (22/58; 37.9%) were in the three largest clusters.

The ages between which transmission occurred varied by setting. When considering all instances of likely, possible and probable transmission, transmission occurred between a relatively wide range of age groups in household settings; 17.0% (9/53) between 15 and 34 year olds (Figure 4A) (Appendix Figure 3). In contrast, among instances of transmission that occurred in social/religious settings for which the age of a source case could be defined, 46.7% (7/15) occurred between 15 and 34 year olds (Figure 4B). If we infer the age of 23 of the 24 cases with uncertain possible source cases by averaging the age of all their possible source cases (inference was impossible for one probable case because they were epidemiologically linked to cases from multiple sub-clusters), 66.7% (26/39) occurred between 15 and 34 year olds and 38.5% (15/39) between 15 and 24 year olds (Figure 4H).

Association between TB case characteristics and *Mtb* transmission

Likely source cases most often transmitted to one other person (18/29; 65.5%, median = 1, interquartile range = 1 −2) (Appendix Figure 4) and there were three instances of likely transmission chains, i.e. a likely secondary case becoming a likely source case themselves, and transmitting onwards.

Likely source cases often reported a cough (24/29; 82.8%), and all but one reported symptoms (28/29; 96.6%) although, inconsistently, the case-notes of the single asymptomatic case mentioned a recent history of an intermittent cough (Table 1).

The multivariable Poisson regression model included the covariates sputum smear positivity, sex, presence of a cough, place of birth (overseas or not), and whether or not the chest x-ray suggested past TB. In this model, the only case characteristic found to be associated with the number of resultant secondary cases, as indicated by likelihood ratio tests with a p value of <0·001, was sputum smear positivity (incidence rate ratio [IRR] 12.74, 95% CI 4.44−36.60, $p < 0.001$) as compared to sputum smear negativity. Cases that reported having had a cough were also associated with a higher number of secondary cases than those without (IRR 5·03, 95% CI 1.60−15.83, $p = 0.006$, likelihood ratio p value = 0.0015). An assessment of the random effects of clusters (≤5 SNPs) (Appendix Figure 5) revealed Lineage 2 and Lineage 4 to make up a higher proportion of the results above the population mean estimate (45.8% and 33.3%, respectively), and they also had higher average random effect terms (5.9 and 22.9, respectively) than Lineages 1 and 3 (2.4 and 2.8, respectively), which may imply higher transmissibility of these lineages (Appendix Figure 5).

The multivariable regression model included the covariates sputum smear positivity, event year, presence of a cough, cavity, age group and whether they had ever spent time in a correctional facility. In this model, the only case characteristic found to be associated with being a likely or possible source case, as indicated by likelihood ratio tests with a p value of <0.001, was the event year. Cases notified in 2020 were less likely to be identified as possible index cases (odds ratio [OR] 0.07, 95% CI 0.02−0.28, p <0.001) than cases in 2017 (Appendix Table 4), which may relate to the increased likelihood of being identified as a possible index case with increasing time in the study. An assessment of the random effects of clusters (≤5 SNPs) (Appendix Figure 6) revealed Lineage 2 and Lineage 4 to make up a higher proportion of the results above one (33.3% each), which may imply higher transmissibility of these lineages, although and their average random effect terms (3.2 and 8.0, respectively) were similar to Lineages 1 and 3 (5.1 and 3.0, respectively).

During the study period, 20−24 year old males made up the highest proportion of all cases who were smear positive (38/297; 12.7%, Appendix Figure 6A), and the groups with the highest probability of being smear positive were 15−19 year olds females (9/30; 30.0%) and, again, 20−24 year old males (38/141; 27.0%, Appendix Figure 6B).

Pulmonary involvement was excluded as a variable in the models because it perfectly predicted transmission, but in a separate multivariable logistic regression model, lineage was associated with pulmonary involvement (likelihood ratio p value < 0.001). Those with *Mtb* strain Lineage 2 (OR 1.80, 95% CI 1.26−2.58, $p = 0.001$) were more likely to have pulmonary involvement, compared to those with Lineage 1, and those with Lineage 3 were less likely (OR 0.66, 95% CI 0.46−0.94, $p = 0.02$) (Appendix Table 5).

Potentially-related genomic clusters (6-12 SNPs)

During the study period, 50 cases clustered to other cases in the 6−12 SNP range, 19 of which did not cluster at ≤5 SNPS to any other cases. Of the 40 potential secondary cases in this group, local transmission was impossible for five (5/40; 12.5%) (Box 1), compared to one (0.7%) of the 140 potential secondary cases in highly related clusters. All but one of the 35 cases of possible local transmission (at 6−12 SNPs) shared the same country of birth, lived in the same local geographical area, or were diagnosed within the same year as their possible source cases.

In one of the largest clusters (≤12 SNPs), which included three highly-related clusters (≤5 SNPs), three cases only had epidemiological links (including one household contact) with cases in a different sub-cluster.

Discussion

Genomic and epidemiological investigations in our low-incidence setting from 2017 to 2020 determined that 6.4% of sequenced isolates from TB cases likely resulted from local transmission, and a further 4.5% were possibly due to local transmission, with most instances occurring beyond household contacts. Most sequenced TB isolates were genotypically distinct, and are likely to reflect TB incursions resulting from infection that was acquired overseas.[39] Australia has seen high levels of immigration from high incidence settings for many years.[40]

Molecular typing can be used to support TB program activities in several ways. One of its uses is to identify unsuspected instances of *Mtb* transmission, thus prompting further epidemiological investigation and possible opportunities for prevention, and the potential advantage of WGS compared to other typing techniques lies in its superior resolution.[14,33] While we could not compare typing techniques head-to-head, a previous report of clustering and transmission in our setting (2003–2010) based on MIRU-VNTR provides an opportunity for comparison. In the earlier study, 17.0% of genotyped isolates were genotypically indistinguishable, compared to 10.8% in our study, likely reflecting the improved discriminatory power of WGS.[41] It is possible that the lower proportion of cases requiring additional epidemiological investigation allows for, or motivates, more comprehensive investigations.[32] In one instance during the study period, investigations prompted by WGS led to the identification of a new transmission site, leading to additional active case finding and use of preventative therapy; MIRU-VNTR typing results may not have facilitated a similar response. Second, genotyping methods are also used to identify cases due to laboratory contamination, allowing for the cessation of unnecessary treatment and opportunities to review laboratory handling practices. A similar proportion of cases with sequenced isolates were identified as resulting from laboratory contamination in both this and the earlier study (0.8% and 0.7%, respectively),[18] suggesting both methods are similarly able to inform on these events. Third, WGS is able to identify possible antimicrobial resistance, and may do so sooner than culture-based methods, allowing for the timelier provision of appropriate treatment regimens.[13] Although it has not yet been used for this purpose in our setting, we found high concordance between phenotypic and genotypic sensitivity of isolates.

More broadly, molecular and epidemiological data can also inform the characterisation of *Mtb* transmission in a setting and, therefore, inform TB control strategies. Regardless of the typing method or transmission classification (i.e. likely or possible), *Mtb* transmission commonly occurred beyond household contacts in our low-incidence setting and these instances accounted for a large proportion of the potentially infectious (i.e. smear positive) secondary cases. We found that young adults had a relatively high likelihood of pulmonary and smear positive disease, and transmission among young adults in social and religious settings accounted for the majority of the identified extra-household instances and the largest clusters, emphasising the need for broad contact tracing investigations, and the possible benefit of directing prevention efforts to certain demographic groups or clusters. However, transmission also remained uncharacterised for many highly-related clusters, which prevents any public health response in our setting, despite genomic evidence. While it is possible that some of these unexpected clusters may relate to transmission that occurred outside geographical boundaries, with frequent interstate and overseas travel reported by many in the study, some are also likely to represent extra-household local transmission not identified during contact tracing investigations. As has been noted in other settings,[32,42] epidemiological investigations were sometimes challenging in our setting due to complex social situations and patient reluctance to share knowledge of contacts, potentially preventing the identification of epidemiological links. Additionally, as observed in our study, *Mtb* transmission can occur during casual contact, but these instances are more challenging to identify, and it may only sometimes be possible to do so retrospectively. Because of their important contribution to transmission, continued systematic, comprehensive, investigations of unexpected clusters will be important to inform our understanding and practices into the future. Linking genomic data between jurisdictions and other countries may also inform our understanding of the significance of WGS clustering in the absence of identified epidemiological links.

It is interesting to consider how well our results may be generalised to other settings, or even to our own setting across time. Genotyping studies in high incidence settings have similarly highlighted the importance of extra-household transmission[43–45] and, consistently, in our setting only 14.0% of overseas born TB cases during the study period reported having household or close contact with a TB case. A WGS study in Malawi estimated that only 9.4% of confirmed TB transmission was from known contacts (using ≤10 SNPs to confirm transmission),[43] lower than we observed among clustered cases in our setting, even if we applied a ≤12 SNP cut-off. It is possible that the relative importance of household versus extra-household transmission in different settings may be influenced by factors such as household size, social mixing patterns, urban density and community TB incidence. Improving our understanding of these sociodemographic influences may usefully advise case-finding efforts, including contact-tracing, particularly in high incidence settings. Extra-household social mixing patterns have been modified

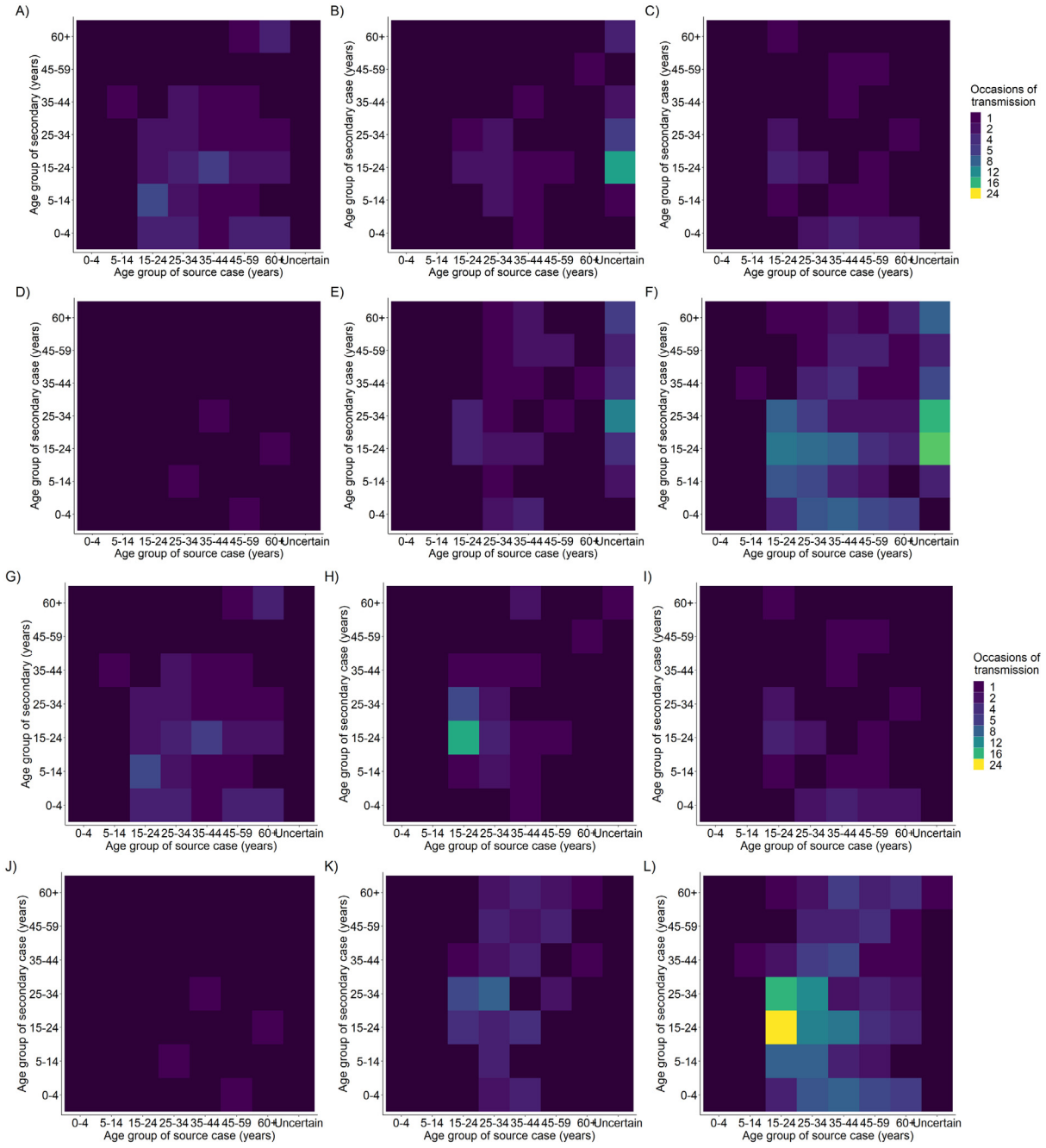

Figure 4. Transmission matrices showing the age groups between which local transmission occurred in Victoria, 2017-2020, for those instances where transmission was likely, possible or probable in A) household, B) social/religious C) family (living in different households), D) other E) unknown, and F) all settings. In Panel 2 the average age of all possible sources cases is used to estimate the age of source cases for 53 secondary cases with multiple possible source cases. The transmission matrices include the age groups of source and secondary cases for instances where the local transmission is: G) household, H) social/religious, I) family (living in different households, J) other, K) unknown, and L) all settings. Ages are those at the time of case diagnosis for both source and secondary cases. "Other" includes instances of transmission in health, education and childcare setting. Inferring the age of the possible source case for one probable secondary case was impossible because they were epidemiologically linked to cases from multiple sub-clusters.

during the COVID-19 pandemic (e.g. due to "stay-at-home" restrictions) and this may provide an opportunity to assess their influence on *Mtb* transmission in our setting and others. To fully assess this in our setting, further collection and analysis of WGS data from 2021 and beyond will be required.

The use of ≤5 SNPs as a primary cut-off to define highly related clusters was largely supported by our study findings. While there were three instances where cases only shared epidemiological links with potentially related source cases (6–12 SNPs), these cases were all part of a large cluster (≤12 SNPs) and may reflect our incomplete understanding of transmission chains, rather than the threshold. However, given this, and given that cluster differences may expand over longer follow up periods due to accumulated mutations,[46] continued monitoring of thresholds greater than 5 SNP may have value in our setting.

Multivariate analysis in our cohort indicated that sputum smear positivity was associated with a higher number of secondary cases, but there was also some evidence that the presence of a cough and lineage were associated. The infectiousness of sputum smear positive cases is well understood,[47,48] but there is little empirical evidence supporting the importance of cough for *Mtb* transmission,[49,50] except for a study by Turner *et al.* 2018 that found a weak association between objectively assessed 24 hour cough frequency and the prevalence immunoreactivity among household contacts (p = 0.022).[51] The possible association we found between cough presence and a higher number of secondary cases in our cohort adds molecular epidemiological evidence to support the possible contribution of cough to transmissibility. Cluster random effects in our analysis also suggested there may be differing transmission dynamics by *Mtb* lineage, as other studies have similarly found,[52–55] particularly Lineage 2.[52] However, in addition to Lineage 2, our results also suggested that Lineage 4 may be associated with a higher number of secondary cases, and this has not been commonly observed. For example, a WGS study in Malawi found Lineage 2 and 3 were more likely to transmit than Lineage 4.[55] Two of the largest clusters in our setting were Lineage 4, and so we cannot rule out the possibility that certain social and/or environmental factors in these clusters contributed to our finding, rather than lineage. A further limitation of our analysis was that we necessarily excluded instances of likely transmission with uncertain source cases. Due to the latency of *Mtb*, assigning source cases in any cluster with a history of more than one infectious individual will always be difficult, despite WGS. Unfortunately, this limits our ability to characterise transmission in larger clusters and the degree to which several individuals contribute to transmission versus fewer, highly-infectious, individuals.

There were several other study limitations. First, we did not assess the cost-effectiveness of WGS compared to prior techniques (MIRU-VNTR), as there was only a short period where both techniques were used. A formal evaluation should include an assessment of the cost-benefit of implementation of *Mtb* WGS, as previously described by our group.[56] Second, the use of typing methods to characterise *Mtb* transmission omits instances involving unsequenced isolates. We included 'probable' transmission in our analysis for this reason but acknowledge that instances without epidemiological links may still have been missed, as will those involving unsequenced isolates beyond the four-year study period or those involving the cases that were unsequenced in late 2017. Therefore, this may have led to an underestimation of clustering in our study, and if the timing of disease progression differs by lineage,[55] this may also have influenced our findings. Additionally we acknowledge that bias is possible when assigning possible source and secondary cases, and may have occurred in our study. For example, the identity of all past pulmonary cases in a cluster cannot be divulged to each new case, and once an epidemiological link has been established between a case and a potential source case in a cluster, further investigation is less likely, despite the possible existence of other links. This may bias source case designations to those that are more easily identified, e.g. household contacts. Finally, TB contacts can be prescribed preventive treatment in our setting, and so it's possible this may have affected the degree and distribution of clustering observed, although because treatment commencement and completion are not systematically monitored in our setting, the impact is uncertain.

The incorporation of WGS into programmatic management continues to evolve. While WGS findings in our setting have enabled the more accurate characterisation of *Mtb* transmission and, in one instance, prompted the need for additional public health investigations, our findings also illustrate that in many cases even after incorporating both genomic and epidemiological data, connection between some reported cases remains uncertain. Existing approaches to contact tracing should be reviewed in light of genomic evidence of transmission, ideally using a systematic approach to implementation.[56] In our setting, we now plan to establish and evaluate national consensus standards for how the insights gained from WGS derived genomic data can best inform and guide programmatic responses in future.

Contributors

Katie Dale contributed to study design, analysis, visualisations and writing. Justin Denholm conceived the study and contributed to study design, supervision and draft development. Ee Laine Tay contributed to data curation and interpretation, study design and analysis. Maria Globan and Simone Bittmann contributed to

data collection, investigation and interpretation. Norelle Sherry and Kristy Horan contributed to data collection, investigation, interpretation and presentation and drafted sections of the methods. Deborah Williamson, Benjamin Howden and Susan Ballard were involved in conceptualisation, supervision and funding acquisition. Niamh Meagher and David Price provided guidance regarding the statistical analyses, including the methods, interpretation and reporting. All authors contributed to article revisions.

Data sharing statement

Deidentified data is available on request in conjunction with a protocol approved by an appropriate human research ethics committee.

Declaration of interests

None to declare.

Acknowledgements

The authors acknowledge the TB nurses at the Victorian TB Program, the specialist TB clinicians at the network of Victorian TB clinics and the scientists at the Victorian Mycobacterium Reference Laboratory and the Microbiological Diagnostic Unit for generating the notification, clinical and laboratory data used in this study. Funding for this work was provided by the Victorian Department of Health.

Ethics committee approval

All data for this project were collected under the Public Health and Wellbeing Act 2008 (Victoria). Ethical approval was received from the University of Melbourne Human Research Ethics Committee (study number 1954615).

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.lanwpc.2022.100556.

References

1 Duffy SC, Srinivasan S, Schilling MA, et al. Reconsidering Mycobacterium bovis as a proxy for zoonotic tuberculosis: a molecular epidemiological surveillance study. *The Lancet Microbe*. 2020;1(2):e66–e73.

2 WHO. Tuberculosis. 2006.

3 Van Soolingen D. Molecular epidemiology of tuberculosis and other mycobacterial infections: main methodologies and achievements. *J Intern Med*. 2001;249(1):1–26.

4 Filliol I, Driscoll JR, Van Soolingen D, et al. Global distribution of Mycobacterium tuberculosis spoligotypes. *Emerg Infect Dis*. 2002;8(11):1347–1349.

5 Shabbeer A, Cowan LS, Ozcaglar C, et al. TB-Lineage: an online tool for classification and analysis of strains of Mycobacterium tuberculosis complex. *Infect Genet Evol*. 2012;12(4):789–797.

6 Fitzgibbon MM, Gibbons N, Roycroft E, et al. A snapshot of genetic lineages of Mycobacterium tuberculosis in Ireland over a two-year period, 2010 and 2011. *Euro Surveill*. 2013;18(3):20367.

7 Dale JW, Brittain D, Cataldi AA, et al. Spacer oligonucleotide typing of bacteria of the Mycobacterium tuberculosis complex: recommendations for standardised nomenclature. *Int J Tuberc Lung Dis*. 2001;5(3):216–219.

8 van Embden JD, van Gorkom T, Kremer K, Jansen R, van Der Zeijst BA, Schouls LM. Genetic variation and evolutionary origin of the direct repeat locus of Mycobacterium tuberculosis complex bacteria. *J Bacteriol*. 2000;182(9):2393–2401.

9 Frothingham R, Meeker-O'Connell WA. Genetic diversity in the Mycobacterium tuberculosis complex based on variable numbers of tandem DNA repeats. *Microbiology*. 1998;144(Pt 5):1189–1196.

10 Supply P, Lesjean S, Savine E, Kremer K, van Soolingen D, Locht C. Automated high-throughput genotyping for study of global epidemiology of Mycobacterium tuberculosis based on mycobacterial interspersed repetitive units. *J Clin Microbiol*. 2001;39(10):3563–3571.

11 Satta G, Lipman M, Smith GP, Arnold C, Kon OM, McHugh TD. Mycobacterium tuberculosis and whole-genome sequencing: how close are we to unleashing its full potential? *Clin Microbiol Infect*. 2018;24(6):604–609.

12 Shea J, Halse TA, Lapierre P, et al. Comprehensive Whole-Genome Sequencing and Reporting of Drug Resistance Profiles on Clinical Cases of Mycobacterium tuberculosis in New York State. *J Clin Microbiol*. 2017;55(6):1871–1882.

13 Walker TM, Kohl TA, Omar SV, et al. Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study. *Lancet Infect Dis*. 2015;15(10):1193–1202.

14 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:S40–S42.

15 Australian Bureau of Statistics. *National, state and territory population: Statistics about the population and components of change (births, deaths, migration) for Australia and its states and territories*. 2022. https://www.abs.gov.au/statistics/people/population/national-state-and-territory-population/latest-release. Accessed 5 May 2022.

16 Australian Bureau of Statistics. *National, state and territory population: Statistics about the population and components of change (births, deaths, migration) for Australia and its states and territories*. 2021. https://www.abs.gov.au/statistics/people/population/national-state-and-territory-population/jun-2021. Accessed 5 May 2022.

17 Victorian Department of Health. *Local Government areas surveillance report*. 2022. https://www.health.vic.gov.au/infectious-diseases/local-government-areas-surveillance-report. Accessed 5 May 2022.

18 Globan M, Lavender C, Leslie D, et al. Molecular epidemiology of tuberculosis in Victoria, Australia, reveals low level of transmission. *The International Journal of Tuberculosis and Lung Disease*. 2016;20(5):652–658.

19 Australian Government Department of Health. *Tuberculosis case definition*. 2010. 22 December 2010; https://www1.health.gov.au/internet/main/publishing.nsf/Content/cda-surveil-nndss-casedefs-cd_tb.htm. Accessed 02 September 2021.

20 Votintseva AA, Pankhurst LJ, Anson LW, et al. Mycobacterial DNA extraction for whole-genome sequencing from early positive liquid (MGIT) cultures. *J Clin Microbiol*. 2015;53(4):1137–1143.

21 Horan K. Troika-TB 0.0.5: A pipeline implementing TB-Profiler for batch detection and reporting of anti-microbial resistance in TB for public health and clinical use. 2020. https://pypi.org/project/Troika-TB/ Accessed 20 April 2022.

22 Wood D. kraken2. 2020. https://github.com/DerrickWood/kraken2/wiki/Manual Accessed 20 April 2022.

23 Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, Hugenholtz P. Selection of representative genomes for 24,706 bacterial and archaeal species clusters provide a complete genome-based taxonomy. *bioRxiv*, 2019. https://doi.org/10.1101/771964.

24 Parks DH, Chuvochina M, Waite DW, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol*. 2018;36(10):996–1004.

25 Lipworth S, Jajou R, de Neeling A, et al. SNP-IT Tool for Identifying Subspecies and Associated Lineages of Mycobacterium tuberculosis Complex. *Emerg Infect Dis*. 2019;25(3):482–488.

26 Phelan JE, O'Sullivan DM, Machado D, et al. Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. *Genome Med*. 2019;11(1):41.

27 Camus J-C, Pryor MJ, Médigue C, ST Cole. Re-annotation of the genome sequence of Mycobacterium tuberculosis H37Rv. *Microbiology*. 2002;148(10):2967–2973.

28 Seeman T. snippy. 2020. https://github.com/tseemann/snippy Accessed 20 April 2022.
29 Minh BQ, Schmidt HA, Chernomor O, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol*. 2020;37(5):1530–1534.
30 Walker TM, Lalor MK, Broda A, et al. Assessment of Mycobacterium tuberculosis transmission in Oxfordshire, UK, 2007-12, with whole pathogen genome sequences: an observational study. *Lancet Respir Med*. 2014;2(4):285–292.
31 Yang C, Luo T, Shen X, et al. Transmission of multidrug-resistant Mycobacterium tuberculosis in Shanghai, China: a retrospective observational study using whole-genome sequencing and epidemiological investigation. *The Lancet Infectious diseases*. 2017;17(3):275–284.
32 Walker TM, Ip CL, Harrell RH, et al. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. *Lancet Infect Dis*. 2013;13(2):137–146.
33 Denholm J, Coulter C, Bastian I, Committee NTA. Defining a tuberculosis cluster or outbreak. *Communicable diseases intelligence quarterly report*. 2016;40(3):E356.
34 R Core Team. In: Computing RFfS, ed. *R: A language and environment for statistical computing*. 2021. editorVienna, Austria.
35 Coll F, McNerney R, Guerra-Assunção JA, et al. A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. *Nature Communications*. 2014;5(1):4812.
36 Lempens P, Meehan CJ, Vandelannoote K, et al. Isoniazid resistance levels of Mycobacterium tuberculosis can largely be predicted by high-confidence resistance-conferring mutations. *Sci Rep*. 2018;8(1):3246.
37 Dale KD, Trauer JM, Dodd PJ, Houben RM, Denholm JT. Estimating long-term tuberculosis reactivation rates in Australian migrants. *Clinical Infectious Diseases*. 2020;70(10):2111–2118.
38 Dale KD, Trauer JM, Dodd PJ, Houben RMGJ, Denholm JT. Estimating the prevalence of latent tuberculosis in a low incidence setting: Australia. *Eur Respir J*. 2018;52(6):1801218. https://doi.org/10.1183/13993003.01218-2018.
39 Jajou R, de Neeling A, van Hunen R, et al. Epidemiological links between tuberculosis cases identified twice as efficiently by whole genome sequencing than conventional molecular typing: A population-based study. *PLoS One*. 2018;13(4):e0195413.
40 Alaridah N, Hallbäck ET, Tängrot J, et al. Transmission dynamics study of tuberculosis isolates with whole genome sequencing in southern Sweden. *Sci Rep*. 2019;9(1):4931.
41 Glynn JR, Guerra-Assuncao JA, Houben RM, et al. Whole Genome Sequencing Shows a Low Proportion of Tuberculosis Disease Is Attributable to Known Close Contacts in Rural Malawi. *PLoS One*. 2015;10(7):e0132840.
42 Verver S, Warren RM, Munch Z, et al. Proportion of tuberculosis transmission that takes place in households in a high-incidence area. *Lancet*. 2004;363(9404):212–214.
43 Auld SC, Shah NS, Mathema B, et al. Extensively drug-resistant tuberculosis in South Africa: genomic evidence supporting transmission in communities. *Eur Respir J*. 2018;52(4):1800246.
44 Meumann EM, Globan M, Fyfe JA, et al. Genome sequence comparisons of serial multi-drug-resistant Mycobacterium tuberculosis isolates over 21 years of infection in a single patient. *Microbial genomics*. 2015;1(5):e000037.
45 Rodrigo T, Caylà JA, García de Olalla P, et al. Characteristics of tuberculosis patients who generate secondary cases. *Int J Tuberc Lung Dis*. 1997;1(4):352–357.
46 Grzybowski S, Barnett GD, Styblo K. Contacts of cases of active pulmonary tuberculosis. *Bull Int Union Tuberc*. 1975;50(1):90–106.
47 Patterson B, Wood R. Is cough really necessary for TB transmission? *Tuberculosis*. 2019;117:31–35.
48 Dowdy DW. Coughing is Not Required to Transmit Mycobacterium tuberculosis: Another Nail in the Coffin. *Am J Respir Crit Care Med*. 2022;206(2):141–143.
49 Turner RD, Birring SS, Darmalingam M, et al. Daily cough frequency in tuberculosis and association with household infection. *Int J Tuberc Lung Dis*. 2018;22(8):863–870.
50 Karmakar M, Trauer JM, Ascher DB, Denholm JT. Hyper transmission of Beijing lineage Mycobacterium tuberculosis: Systematic review and meta-analysis. *J Infect*. 2019;79(6):572–581.
51 Albanna AS, Reed MB, Kotar KV, et al. Reduced transmissibility of East African Indian strains of Mycobacterium tuberculosis. *PLoS One*. 2011;6(9):e25075.
52 Holt KE, McAdam P, Thai PVK, et al. Frequent transmission of the Mycobacterium tuberculosis Beijing lineage and positive selection for the EsxW Beijing variant in Vietnam. *Nat Genet*. 2018;50:849.
53 Guerra-Assunção JA, Crampin AC, Houben RM, et al. Large-scale whole genome sequencing of M. tuberculosis provides insights into transmission in a high prevalence area. *eLife*. 2015;4:e05166.
54 Ferdinand AS, Kelaher M, Lane CR, et al. An implementation science approach to evaluating pathogen whole genome sequencing in public health. *Genome Med*. 2021;13(1):121.
55 Marais BJ, Gie RP, Schaaf HS, et al. The natural history of childhood intra-thoracic tuberculosis: a critical review of literature from the pre-chemotherapy era. *Int J Tuberc Lung Dis*. 2004;8(4):392–402.
56 Seddon JA, Chiang SS, Esmail H, Coussens AK. The Wonder Years: What Can Primary School Children Teach Us About Immunity to Mycobacterium tuberculosis? *Front Immunol*. 2018;9:2946.