MICROBIAL GENOMICS

RESEARCH ARTICLE

Bainomugisa *et al.*, *Microbial Genomics* 2021;7:000573

DOI 10.1099/mgen.0.000573

MICROBIOLOGY SOCIETY

OPEN DATA OPEN ACCESS

Genomic epidemiology of tuberculosis in eastern Malaysia: insights for strengthening public health responses

Arnold Bainomugisa[1], Ella M. Meumann[2,3], Giri Shan Rajahram[4,5,6], Rick Twee-Hee Ong[7], Lachlan Coin[8], Dawn Carmel Paul[9], Timothy William[4,6,10], Christopher Coulter[1] and Anna P. Ralph[1,2,3,*]

Abstract

Tuberculosis is a leading public health priority in eastern Malaysia. Knowledge of the genomic epidemiology of tuberculosis can help tailor public health interventions. Our aims were to determine tuberculosis genomic epidemiology and characterize resistance mutations in the ethnically diverse city of Kota Kinabalu, Sabah, located at the nexus of Malaysia, Indonesia, Philippines and Brunei. We used an archive of prospectively collected *Mycobacterium tuberculosis* samples paired with epidemiological data. We collected sputum and demographic data from consecutive consenting outpatients with pulmonary tuberculosis at the largest tuberculosis clinic from 2012 to 2014, and selected samples from tuberculosis inpatients from the tertiary referral centre during 2012–2014 and 2016–2017. Two hundred and eight *M. tuberculosis* sequences were available for analysis, representing 8% of cases notified during the study periods. Whole-genome phylogenetic analysis demonstrated that most strains were lineage 1 (195/208, 93.8%), with the remainder being lineages 2 (8/208, 3.8%) or 4 (5/208, 2.4%). Lineages or sub-lineages were not associated with patient ethnicity. The lineage 1 strains were diverse, with sub-lineage 1.2.1 being dominant (192, 98%). Lineage 1.2.1.3 isolates were geographically most widely distributed. The greatest diversity occurred in a border town sub-district. The time to the most recent common ancestor for the three major lineage 1.2.1 clades was estimated to be the year 1966 (95% HPD 1948–1976). An association was found between failure of culture conversion by week 8 of treatment and infection with lineage 2 (4/6, 67%) compared with lineage 1 strains (4/83, 5%) (P<0.001), supporting evidence of greater virulence of lineage 2 strains. Eleven potential transmission clusters (SNP difference ≤12) were identified; at least five included people living in different sub-districts. Some linked cases spanned the whole 4-year study period. One cluster involved a multidrug-resistant tuberculosis strain matching a drug-susceptible strain from 3 years earlier. Drug resistance mutations were uncommon, but revealed one phenotype–genotype mismatch in a genotypically multidrug-resistant isolate, and rare nonsense mutations within the *katG* gene in two isolates. Consistent with the regionally mobile population, *M. tuberculosis* strains in Kota Kinabalu were diverse, although several lineage 1 strains dominated and were locally well established. Transmission clusters – uncommonly identified, likely attributable to incomplete sampling – showed clustering occurring across the community, not confined to households or sub-districts. The findings indicate that public health priorities should include active case finding and early institution of tuberculosis management in mobile populations, while there is a need to upscale effective contact investigation beyond households to include other contacts within social networks.

Received 28 July 2020; Accepted 29 March 2021; Published 04 May 2021

Author affiliations: [1]Queensland Mycobacterium Reference Laboratory, Brisbane, Australia; [2]Global and Tropical Health Division, Menzies School of Health Research, Charles Darwin University, Northern Territory, Australia; [3]Department of Medicine, Royal Darwin Hospital, Northern Territory, Australia; [4]Infectious Diseases Society Sabah-Menzies School of Health Research Clinical Research Unit, Kota Kinabalu, Sabah, Malaysia; [5]Department of Medicine, Queen Elizabeth Hospital, Kota Kinabalu, Sabah, Malaysia; [6]Clinical Research Centre, Queen Elizabeth Hospital, Sabah, Malaysia; [7]Saw Swee Hock School of Public Health, National University of Singapore and National University Health System, Singapore; [8]Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia; [9]Kota Kinabalu Public Health Laboratory, Sabah, Malaysia; [10]Gleneagles Hospital Kota Kinabalu, Sabah, Malaysia.

***Correspondence:** Anna P. Ralph, anna.ralph@menzies.edu.au

Keywords: *Mycobacterium tuberculosis*; whole-genome sequencing; Malaysia.

Abbreviations: DNA, deoxyribosenucleic acid; ESS, effective sample size; GATK, genome analysis toolkit; GTR, general time-reversible; HIV, human immunodeficiency virus; HKY, Hasegawa–Kishino–Yano; HPD, highest posterior density; HREC, Human Research Ethics Committee; MCMC, Markov chain Monte Carlo; MDR, multi drug-resistant; MGIT, Mycobacterium Growth Indicator Tube; MRCA, most recent common ancestor; n, number; NCBI, National Center for BiotechnologyInformation; NMRR, National Medical Research Register (Malaysia); RFLP, restriction fragment length polymorphism; SNP, single nucleotide polymorphism; TB, tuberculosis; WGS, whole genome sequencing; XDR, extensively drug resistant.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Six supplementary tables are available with the online version of this article.

DATA SUMMARY

Data for this project are available in the National Center for Biotechnology Information (NCBI) sequence read archive (PRJNA639216, www.ncbi.nlm.nih.gov/bioproject/639216) and Table S4.

INTRODUCTION

Tuberculosis (TB) remains a leading global cause of morbidity and mortality, and the majority of the global burden of TB is in the Asia-Pacific region [1]. In Malaysia, significant overall progress in TB control has occurred, but Sabah state in eastern Malaysia on the island of Borneo is overrepresented in its contribution to the national TB burden [2, 3]. The case notification rate for Sabah is estimated at 144–217 per 100000 population [2]. Sabah is located at a nexus of four countries in close proximity (Malaysia, Indonesia, Philippines, Brunei) and border crossings are common. Like other middle- or high-income settings, a significant proportion of TB cases in Malaysia occur in international migrants. Undocumented migrants face major barriers to health care and the consequences of this include late TB presentations with extensive pulmonary disease, meaning poorer outcomes for affected individuals, and high opportunity for transmission prior to starting treatment [4, 5]. HIV coinfection has been found to be relatively low (<2%) among outpatients in Kota Kinabalu [5]. Drug-resistant TB is not common, although cases of multidrug-resistant (MDR) TB, including extensively drug-resistant (XDR) TB, are observed, especially among hospitalized patients [4, 6].

Tailoring public health interventions to improve TB control in this context requires clearer knowledge of TB transmission patterns. A study in 1999 using a low-discriminatory form of molecular typing (IS*6110*-based restriction fragment length polymorphism) on 49 isolates from east Malaysia and 380 from peninsular Malaysia identified that TB strains with a distinct RFLP pattern were common in the east, in contrast to peninsular Malaysia, where Beijing family strains dominated [7]. To our knowledge, there have been no subsequent studies applying molecular typing or whole-genome sequencing (WGS) to TB in this region. In Kota Kinabalu, comprehensive knowledge of drug resistance patterns has been lacking.

The aim of this study was to determine the proportion of incident TB attributable to recent transmission and the role of migration in driving incident TB in Kota Kinabalu using an archive of prospectively collected *M. tuberculosis* isolates from patients in focal catchment areas, paired with epidemiological data. Based on epidemiologic observations, we hypothesized that there would be multiple TB strains circulating, with evidence of clustering in geographical hot spots. We also investigated the clinical characteristics of TB caused by different lineages, and characterized mutations associated with drug resistance.

METHODS

Setting

This study was undertaken in Kota Kinabalu, Sabah, Malaysia. Kota Kinabalu is the state capital with a population of approximately 650000 and 700–800 TB cases notified per year. Participants were enrolled at Luyang Tuberculosis Outpatient Clinic, then the largest TB clinic in Kota Kinabalu, at that time seeing approximately 200 patients per year, or from Queen Elizabeth Hospital, the only public tertiary referral centre in Kota Kinabalu. The hospital catchment area includes Luyang district and other regional sub-districts.

Impact Statement

Tuberculosis, caused by the bacterium *Mycobacterium tuberculosis*, is a leading priority in eastern Malaysia. This study provides the most comprehensive description of the types of tuberculosis strains circulating in Sabah, eastern Malaysia. We collected sputum samples from people with tuberculosis in 2014–2017. One third were migrants to Malaysia, from the Philippines or Indonesia. *M.tuberculosis* from sputum was analysed using whole-genome sequencing, a method that reveals the exact genetic code. Two hundred and eight *M.tuberculosis* sequences were included. We found numerous strains circulating, most in the lineage 1 family, and some in lineages 2 and 4. People with lineage 2 took longer to clear infection. Eleven small clusters were found where people shared a TB strain, meaning that it was transmitted between them or that they were infected from a common source. Although tuberculosis control focuses on households, at least five clusters affected people from different addresses, probably infected in the community or at work. Some clusters spanned the whole 4-year study period. In one example, a tuberculosis strain was initially sensitive to antibiotics, but was transmitted and evolved into the multi-resistant form. The findings show that tuberculosis prevention, early diagnosis and treatment need to be escalated beyond households, with increasing focus on mobile populations.

Design and participants

M. tuberculosis samples were collected in the following three sub-studies.

Study 1a

One hundred and seventy-two samples were obtained from 4 July 2012 to 1 July 2014 from all consecutive eligible, consenting pulmonary TB patients attending Luyang outpatient clinic as part of a larger study [4, 5, 8, 9]. Participants were eligible if they provided consent, had sputum smear-positive pulmonary TB, were aged ≥15 years, were not pregnant, and had received <7 days' TB treatment. Data included geographical location, illness duration, ethnicity, country of birth (since some ethnicities such as Bajau occur across national borders), number of years residing in Malaysia if not born in Malaysia, history of past TB, HIV status, co-morbidities and clinical and radiological severity measures. Sputum samples at treatment commencement and at 8 weeks were transported to an international reference laboratory for culture and drug susceptibility testing (Singapore General Hospital Laboratory, Singapore). The BACTEC Mycobacterium Growth Indicator Tube (MGIT) 960 tube system was employed for culture. Drug susceptibility testing was performed using the non-radiometric MGIT system for isoniazid, rifampicin, ethambutol, streptomycin and, in the instance of any first-line resistance, also for ofloxacin, kanamycin and ethionamide.

Study 1b

Seven samples were obtained over the same period as in study 1a (4 July 2012–1 July 2014). Clinicians at Queen Elizabeth Hospital were able to utilize the reference laboratory (Singapore General Hospital Laboratory) to obtain mycobacterial culture and susceptibility testing for clinical purposes. No inclusion criteria were applied apart from suspicion of mycobacterial infection and difficulty obtaining results locally. Samples were sputum (*n*=6) or extrapulmonary (*n*=1: psoas abscess pus sample). Resulting mycobacterial isolates were stored. Minimal demographic data were collected.

Study 2

Fifty-five samples were obtained during 18 February 2016–10 April 2017, GeneXpert testing for TB (MTB/RIF) was introduced at the research laboratory at Queen Elizabeth Hospital. Residual samples that had tested positive for *M. tuberculosis* on GeneXpert (*n*=75) were dispatched to the National University of Singapore research laboratory for culture, susceptibility testing and WGS; *M. tuberculosis* sequence data were available from 55 samples. Minimal patient demographic data were collected.

Ethics and consent

Eligible outpatients for study 1a provided written, informed consent. Consent was obtained from a parent/guardian for participants aged 15–17. The need for consent was waived for inpatients (studies 1b and 2) who had samples and data collected as part of clinical care. Ethical approval was obtained from the Medical Research Ethics Committee, Malaysian Ministry of Health (NMRR-11-1051-10491), and the Human Research Ethics Committee of the Northern Territory Department of Health and Menzies School of Health Research, Australia (HREC-2010–1398).

DNA extraction and sequencing

DNA extraction was performed as previously described [10, 11] using mechanical and chemical methods. Library preparation was done using Nextera XT and 100 base pair (bp) paired-end sequencing was undertaken using the Illumina HiSeq 2000 (Illumina, San Diego, CA, USA) at Forensic Scientific Services (Queensland Health, Coopers Plains, Brisbane, Australia). Isolates in study 2 were cultured at the National University of Singapore. Library preparation was done using the NEBNext Ultra DNA Library Prep kit, and 150 bp paired-end sequencing was done using the Illumina HiSeq4000 (Illumina, San Diego, CA, USA) at the Genome Institute of Singapore. Sequencing data for this project are available in the NCBI sequence read archive (PRJNA639216, www.ncbi.nlm.nih.gov/bioproject/639216).

Genome alignment and variant calling

Genomic sequence data from all sub-studies were analysed together. Raw paired-end reads were filtered for length and trimmed for quality using Trimmomatic version 0.27 [12]. The reads were mapped to *M. tuberculosis H37Rv* (GenBank ID: NC000962.3) using Burrows-Wheeler Alignment [13]. Alignments were refined using samtools [14] and Genome Analysis Toolkit (GATK) tools [15] with regard to base recalibration and realignment for possible artefacts. Single-nucleotide polymorphisms (SNPs) and small insertions and deletions (indels) were called using GATK UnifiedGenotyper and annotated using SNPEff [16]. Variants with a minimum read depth of 10 (5 reads in both forward and reverse orientation), with a Phred score >30, <0.6 strand bias and >75% allele frequency were utilized for downstream analysis. Variants within repetitive gene regions such as PPE/PE and consecutive variants within a 10 bp window flanking indels were excluded.

As in previous studies [17–19], SNPs and small indels in resistance-associated genes for first-line and second-line drugs were investigated. The annotated variant file for each isolate was assessed for the presence of mutations in all known genes, including regulatory genes, that are known to confer resistance to TB drugs (Table S1, available in the online version of this article). The identified mutations were further evaluated by using Integrative Genomic Viewer to view the quality of read mapping at the genomic position of the mutation for each isolate. The presence of each mutation was also screened for in the ReSeqTB drug resistance database [20]. When no mutation was detected in the relevant target genes, the isolate was considered to be genotypically susceptible. *M. tuberculosis* lineage SNP typing was performed according to Coll *et al.* [21]. SNPs specific to each sub-lineage clade were compiled (Table S2).

Phylogenetic analysis

A maximum-likelihood phylogenetic analysis of the lineage 1 study genomes and publicly available *M. tuberculosis* genomes from one other study [22] was undertaken using RAxML v7.4.2 [23]. A general time-reversible (GTR) model with rate heterogeneity accommodated by using discrete rate categories was used with 1000 bootstrap replicates, and trees were visualized using FigTree version 1.4.2 (http://tree.bio.ed.ac.uk/software/figtree) [22].

Molecular dating of sub-lineage 1.2 isolates using BEAST 1.8.2 [24] was performed as previously described [10]. We assessed the temporal signal in a neighbour-joined tree (K80 substitution model) by undertaking regression of the root-to-tip distance over time using TempEst v1.5. Although there was a linear correlation between sampling time and root-to-tip distance, the temporal signal (R^2=0.012) was not strong.

We then used Bayesian evaluation of temporal signal to further evaluate the molecular clock [25, 26]. We ran analyses with strict and relaxed molecular clocks with and without the tip dates and found that the log Bayes factor supported the strict clock (Table S3). For these analyses we used an informative uniform prior distribution on substitution rate (initial value 7.9×10^{-8} nucleotide changes/site/year) as previously reported [27, 28], a coalescent constant size demographic model, the Hasegawa–Kishino–Yano (HKY) substitution model with four gamma categories, and Markov chain Monte Carlo (MCMC) chain length of 70 million (10% burn-in) with sampling of every 10000 traces/trees. For marginal likelihood estimation using path sampling (PS) and stepping stone sampling (SS), we ran MCMC chains for at least 10 million and 100 path steps for every 1000 iterations.

We then ran different population demographic models (constant, exponential, expansion and Bayesian skyline) with the strict clock, and used the same parameters described above for rate model among sites and MCMC chains. We also ran the parameters as above using a (GTR) substitution model in place of the HKY model. We calculated Bayes factors from marginal likelihood estimates obtained from path sampling and stepping stone sampling [29]. The final model with the best Bayes factor included a strict molecular clock, an expansion population demographic model and the GTR substitution model (Table S4), and was run three times to confirm consistency. Tracer v1.6 was used to examine log files for each model's convergence, chain length and effective sample size (ESS >200), which suggested adequate independent draws from the posterior sample for sufficient statistical support. TreeAnnotator was used to obtain the best supported topology under the maximum clade credibility method.

Transmission cluster analysis

With exclusion of SNPs in known drug resistance genes, pairwise SNP differences were computed using the Ape package in R [30]. Transmission clusters among the different lineages were identified using TransCluster [31] at a transmission threshold (T) of 19 and transmission rate (β) of 2. Rather than using SNP distance thresholds alone to define clusters, this method takes into account the elapsed time between cases and an estimated transmission rate in combination with SNP distances to estimate the number of intermediate transmissions between cases. The transmission threshold is the threshold under which the estimated number of intermediate transmissions must be in order for two cases to be defined as clustered. The transmission rate is the rate at which intermediate cases occur in the total time elapsed between the most recent common ancestor of two sampled hosts and sampling events.

Statistical analyses

Statistical analyses were performed in R [32]. Differences in patient characteristics between the lineages/major sub-lineages were assessed using Fisher's exact test. Univariate and multivariable regression analyses were computed to evaluate the relationship between patient characteristics and cavity as an outcome variable.

RESULTS

Study population and characteristics of TB patients

We conducted a prospective genomic epidemiology study in Kota Kinabalu in eastern Malaysia from 2012 to 2014 in outpatients (n=172) (study 1a) and accessed selected samples from inpatients (n=7) during the same period (study 1b), providing 179 mycobacterial isolates (Fig. 1). After exclusions of contaminated or culture-negative samples and non-tuberculous mycobacteria, 162 samples (155 outpatients, 7 inpatients) from 154 individuals were available for analyses. This represents 10% of all 1487 TB cases registered in Kota Kinabalu during the same period. Geographical location information and ethnicity were available for all outpatient isolates.

We also conducted a prospective diagnostic study of inpatients from 2016 to 2017 in Kota Kinabalu (study 2), providing 55 *M. tuberculosis* isolates (54 with usable sequence data). This represented 6% of the 977 registered TB cases in Kota Kinabalu during that period. In total, 216 *M. tuberculosis* (208 non-duplicates) sequences were available for analysis.

MTBC population structure

The majority of strains were lineage 1 (Indo–Oceanic) (195/208, 94%) followed by lineage 2 (East Asian) (8/208, 4%) and lineage 4 (Euro–American) (5/208, 2%). All lineage 2 isolates were in the 2.2 sub-lineage, and all lineage 4 strains were in the 4.3.4.1 sub-lineage (Table S6). The lineage 1 strains were diverse with dominance of sub-lineage 1.2.1 (192/195, 98%), and single isolates for sub-lineages 1.1.1, 1.1.2, 1.1.3 and 1.2.2.

Phylogenetic comparison of lineage 1 strains with a collection of 480 lineage 1 genomes from Northern Thailand [22] revealed the sub-division of 1.2.1 into four major clades; 1.2.1.1 (n=16, 8.3%), 1.2.1.2 (n=52, 27.1%), 1.2.1.3 (n=122, 63.5%) and 1.2.1.X (n=2, 1%) (Fig. 2). While 1.2.1.X appears to be a new clade, further studies in this setting with more samples will be required to confirm this. Strains from Kota Kinabalu had limited interspersion with the Thai genomes.

The lineage 1.2.1.3 isolates were most widely distributed, being found in 18/20 sub-districts (Fig. 3). The cross-roads sub-district (Penampang) had the greatest diversity of lineages and sub-lineages.

Characteristics of Kota Kinabalu patients

Patients were mostly smear-positive (164 out of 196 with known smear status, 84%), consistent with the inclusion criteria of study 1a, male (90/153, 59%) and HIV-negative (133/135, 99%) (Table 1). One third of the patients (52/152, 34%) were born outside Malaysia. We examined associations between clinical characteristics, available for study

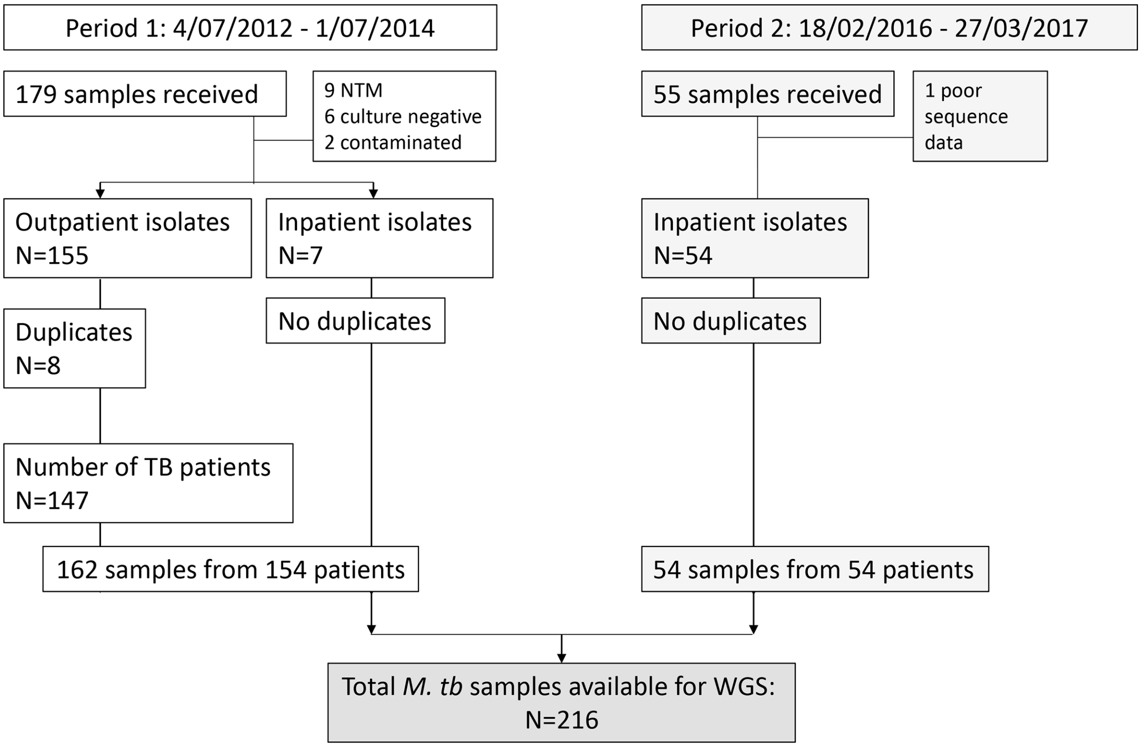
Period 1: 4/07/2012 - 1/07/2014
179 samples received
9 NTM
6 culture negative
2 contaminated
Outpatient isolates N=155
Inpatient isolates N=7
Duplicates N=8
No duplicates
Number of TB patients N=147
162 samples from 154 patients
Period 2: 18/02/2016 - 27/03/2017
55 samples received
1 poor sequence data
Inpatient isolates N=54
No duplicates
54 samples from 54 patients
Total *M. tb* samples available for WGS: N=216

Fig. 1. Study diagram. Figure illustrates numbers of samples provided for analysis from patients in periods 1 and 2; numbers of samples obtained from the different patient cohorts (outpatients and inpatients); final numbers available for analysis. NTM, non-tuberculous mycobacteria.

1a patients, and genotype. Patients infected with lineage 2 strains were slightly more likely to have cavitary disease [5/6 (83%)] than those with non-lineage 2 strains [81/118 (69%)], and to have had a longer period of illness prior to treatment commencement [median 11 weeks (range 3–24)] compared with other lineages [8 weeks (range 1–52)], but these findings were not statistically significant. Persistent *M. tuberculosis* culture positivity at 8 weeks, assessed in study 1a, was uncommon overall (Table 1), but appeared to be more common with lineage 2 (4/6, 67%) than lineage 1 strains (4/83, 5%) (P<0.001). This was not accounted for by drug resistance profile – all those with lineage 2 in study 1a had fully susceptible TB. Within lineage 1 sub-lineages, there were no associations with patient age, ethnicity or other characteristics (Table S5).

Molecular clock phylogeny

We estimated the molecular clock phylogeny and mutation rate of lineage 1 isolates using a previously described method [10, 19]. The mutation rate was estimated to be 0.56 SNPs/genome/year [95% highest posterior density (HPD) 0.23–0.79], similar to other published studies [27, 33, 34]. The three major clades (1.2.1.1, 1.2.1.2 and 1.2.1.3) were estimated to have had a time to most recent common ancestor (MRCA) of 1966 (95% HPD 1948–1976), while clade 1.2.1.X diverged much earlier, rooting the tree, although sample numbers were low. There was no phylogenetic clustering of drug-resistant strains (Fig. 4).

Transmission links

Eleven pairs or small case clusters were identified, each comprising two to five individuals (Fig. 5). Two clusters occurred within single households, five clusters included people living in different sub-districts and there was insufficient epidemiologic information for four clusters. Seven of 22 people involved in clusters with known ethnicity were international migrants. Two clusters only affected migrants and one cluster involved migrants and Malaysian-born individuals. Five clusters included people recruited during different study periods [for example, study 1a (2012–2014) and study 2 (2016–2017)]. The median duration between TB diagnosis of the first and last cases in a cluster was 19 months (range 0–49 months). Of lineage 1 strains, 21/196 (11%) clustered, compared with 4/12 (33%) lineage 2 and 2/5 (40%) lineage 4 strains. One pair involved a drug-susceptible isolate collected from an outpatient's sputum sample on 25 October 2013, and a multi-drug-resistant isolate (resistant to rifampicin and isoniazid) collected from another individual on 6 June 2016 during hospitalization for retreatment of TB (Fig. 5, Table S6, samples OP139 and IP3-2), indicating development of acquired resistance during treatment for an initially drug-susceptible strain of TB.

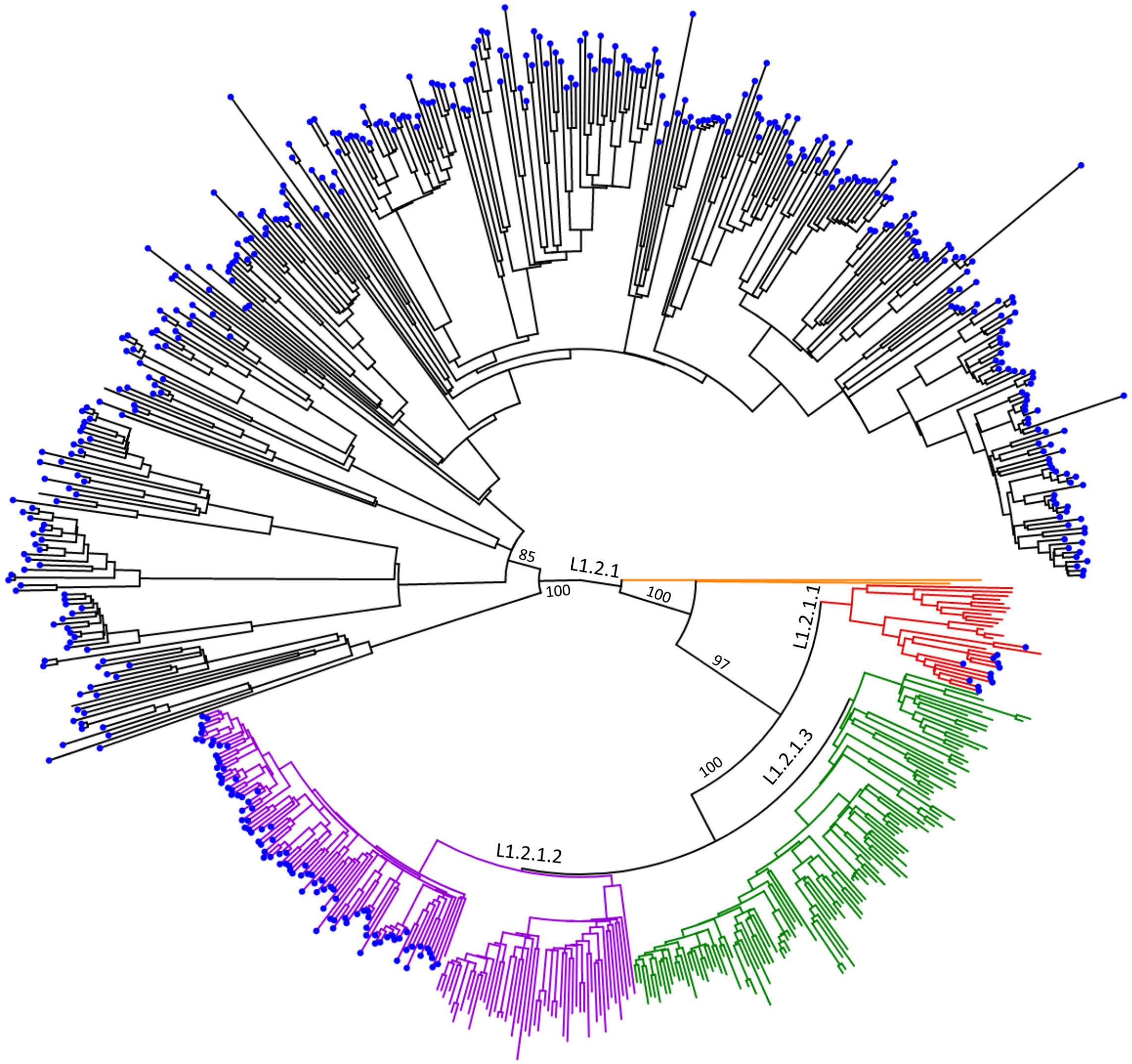

Fig. 2. Phylogeny of lineage 1 isolates from Kinabalu, Malaysia and 480 publicly available Thai *M. tuberculosis* genomes [22] (blue tips). Maximum-likelihood tree with bootstrap values above 75 for all major branches created using RAxML at 1000 bootstraps, GTRCAT nucleotide substitution model based on 8321 SNPs. Clade branch colours: 1.2.1.1 – red, 1.2.1.2 – purple, 1.2.1.3 – green, 1.2.1.X – yellow.

Phenotypic and genotypic resistance

Of the 208 sequenced isolates from unique patients, 193 (93%) were phenotypically sensitive to all first-line agents. Three inpatients had MDR TB isolates; no XDR TB was detected (Table S6). One phenotypically rifampicin-susceptible isolate from an inpatient was genotypically MDR, since it had *rpoB* (p.Leu452Pro) and *katG* (p.Ser315Thr) mutations, plus an additional *embB* (p.Tyr319Asp) mutation. One MDR isolate had a rare nonsense point mutation (p.Glu334X) within the *katG* gene (Fig. S1). Two MDR isolates had *katG* mutations (p.Ser315Thr and p.Gly156Asp). Mono-resistance to isoniazid was identified among eight (3.7%) isolates; five had a *fabG1-inhA* promoter mutation (C-15T) and one had a rare nonsense point mutation (p.Gln50X) within *katG*. Rifampicin mono-resistance in two isolates was due to *rpoB* p.Ser450Leu and p.Ser441Leu mutations. Three streptomycin-resistant isolates were identified, with diverse genetic causes comprising mutations in *gidB* (p.Leu49Arg and p.Glu120fsX), *rrs* (A908G and A907G) and *rpsL* (p.Lys88Arg), respectively.

DISCUSSION

In this most comprehensive description of the genomic epidemiology of tuberculosis in eastern Malaysia to date, we have shown the dominance of lineage 1 *M. tuberculosis* with high strain diversity, and clustering across time and place consistent with frequent introductions. In this multicultural setting, lineage 1 was common across the three main ethnic groups represented (Malaysian, Filipino, Indonesian), with no observable association between lineages or sub-lineages and ethnicity. This concurs with the known high mobility across regional boarders. Temporal reconstruction demonstrated

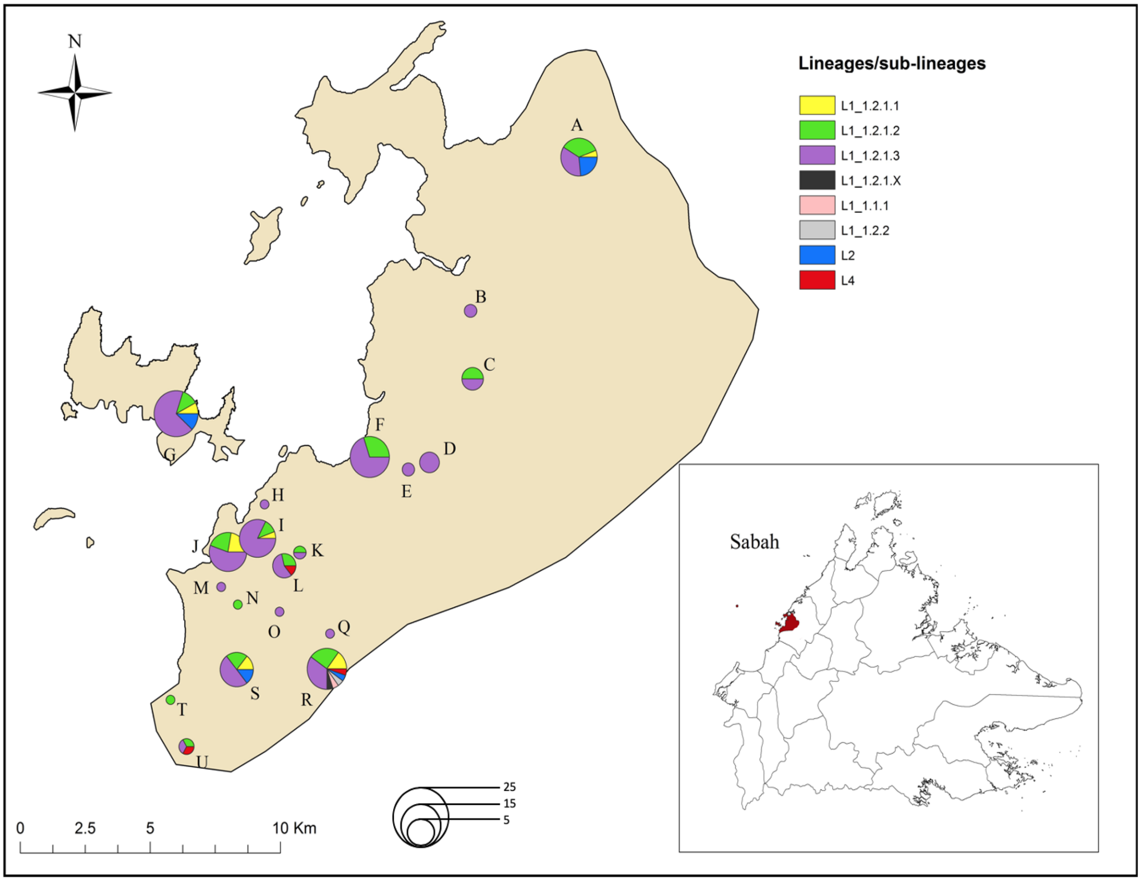

Fig. 3. Map of Kota Kinabalu highlighting the geographical distribution of *M. tuberculosis* cases according to lineages/sub-lineages. Letters represent the different sub-districts/villages: A, Telipok; B, Sulaman; C, Menggatal; D, Inanam; E, Kolombong; F, Likas; G, Gaya; H, Kinarut; I, Kota Kinabalu; J, Sembulan; K, Bundit Padang; L, Luyang; M, Tanjung Aru; N, Kopungit; O, Lido; Q, Bundusan; R, Taman Penampang; S, Kepayan; T, Petagas; U, Putatan. Inset: geographical location of the eastern part of Kota Kinabalu district (red) within Sabah state, Malaysia.

that the three clades (L1.2.1.1, L1.2.1.2, L1.2.1.3) of the dominant sub-lineage L1.2.1 have been spreading for over four decades, while there is some evidence of the ancestral strain still being in circulation. This insight into the emerging population structure of strains circulating provides a reference basis for future understanding of the evolutionary history of strains at a larger scale.

In contrast with our hypothesis, we found less evidence of clustering than expected. This is likely to be attributable to incomplete sampling. Our sampling strategy resulted in the inclusion of only a fraction of all notified cases in Kota Kinabalu during the study periods; however, the sampling at the outpatient clinic in study 1a was intensive, enrolling every sequential consenting smear-positive patient aged >15 years living in the catchment area of that clinic. Given this approach, we expected to find more linkage of cases within families or within the catchment area. Instead, the data suggest wide transmission networks in communities of diverse circulating TB strains. In addition to incomplete sampling, cases missing from this analysis include smear negative and extrapulmonary cases (in whom sample collection is challenging and diagnosis is often on clinical grounds) [35] and undiagnosed TB cases. Incomplete ascertainment of cases is recognized as a problem in all high-TB-burden settings.

An alternative potential explanation for limited evidence of household clustering is that contact tracing at the household level may be effectively limiting household-based clusters. Qualitative work has demonstrated that contacts identified through school or workplace screening programmes in Kota Kinabalu are less likely to attend the clinic for screening than household contacts [36]. Together with the current data in which linked cases spread beyond households or sub-districts, this emphasizes the need for contact investigation to be better implemented in schools, workplaces and social networks. The example of drug-susceptible TB being transmitted and acquiring resistance (Fig. 5; airing on OP139 and IP3-2) highlights the need for effective contract tracing to prevent multidrug-resistant TB. Also, none of the people identified in TB clusters in this study were HIV-positive or children aged <5 years who, according to national [37] and international guidelines at the time of the study [38], are the only groups offered treatment for TB infection (latent TB).

Table 1. Patient demographic and clinical profiles

Factor	Category	Total (%)	Lineage 1	Lineage 2	Lineage 4	*P*-value
Admission status	**Inpatient**	61 (28.6)	57	2	2	
	Outpatient	152 (71.4)	139	11	2	**0.338**
Sputum microscopy	**Positive**	164 (83.7)	148	13	3	
	Negative	32 (16.3)	31	0	1	**0.198**
	ND	17	17	0	0	
No. of isolates involved in genomic clusters	**Yes**	27 (12.7)	21	4	2	**0.013**
	No	186 (87.3)	175	9	2	
Age (years)	**<18**	12 (8)	11	1	0	**0.486**
	19–29	58 (38.7)	54	2	2	
	30–39	26 (17.3)	23	3	0	
	>40	54 (36)	49	5	0	
	ND	**63**	**59**	**2**	**2**	
Sex	**Female**	63 (41.2)	58	3	2	**0.672**
	Male	90 (58.8)	80	8	2	
	ND	60	58	2	0	
Body mass index (kg m^{-2})	**<18**	75 (55.6)	64	10	1	**0.166**
	19–24	51 (38.6)	50	1	0	
	25–30	6 (4.5)	6	0	0	
	>30	1 (0.8)	1	0	0	
	Mean	18.5				
	ND	80	75	2	3	
Culture+ at 8 weeks	**Yes**	8 (8.8)	4	4	0	**<0.001**
	No	82 (91.1)	78	3	0	
	ND	118	114	2	4	
HIV status	**Positive**	2 (1.5)	2	0	0	**0.995**
	Negative	133 (98.5)	121	11	1	
	ND	78	73	2	3	
Haemoptysis	**Yes**	35 (26.1)	30	5	0	**0.239**
	No	99 (73.9)	92	6	1	
	ND	79	74	2	3	
Smoking	**Yes**	69 (49.3)	63	6	0	**0.884**
	No	71 (50.7)	65	5	1	
	ND	73	68	2	3	
Presence of cavity	**Yes**	51 (52.6)	42	8	1	**0.414**
	No	46 (47.4)	42	3	1	
	ND	116	112	2	2	

Continued

Table 1. Continued

Factor	Category	Total (%)	Lineage 1	Lineage 2	Lineage 4	*P*-value
Country of birth	**Malaysia**	100 (65.8)	91	7	2	**0.568**
	Indonesia	6 (3.9)	5	1	0	
	Philippines	46 (30.3)	43	3	0	
	ND	61	57	2	2	

ND, not determined; *P*-values calculated using Fisher's exact test.

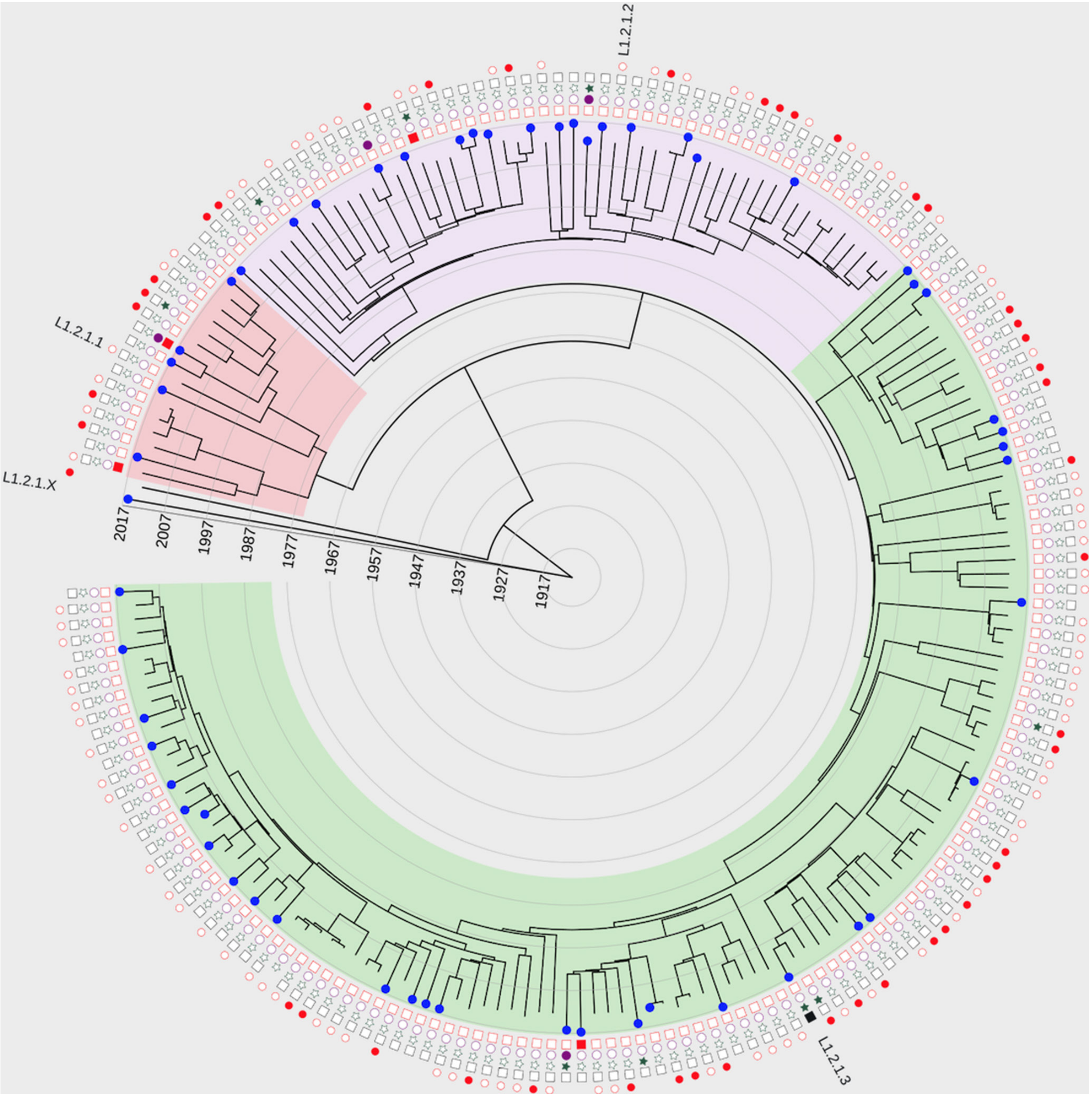

Fig. 4. Bayesian timed phylogeny of the dominant sub-lineage 1.2.1 and drug susceptibility profiles. Tree branch ends: blue, inpatients; without colour, outpatients. Field shapes on the outer edge of the tree: red square, streptomycin-resistant; purple circle, rifampicin-resistant; green star, isoniazid-resistant; black square, ethambutol-resistant. Outermost circle shows immigrant status: red circle, immigrant; open circle, Malaysian born; missing circle, unknown. Clades: green, L1.2.1.3; purple, L1.2.1.2; red, L1.2.1.1; white, L1.2.1.X. L1.2.1.1 had an MRCA of 1966 (95% HPD 1948–1976); L1.2.1.X was rooted in 1910 (95% HPD 1888–1930).

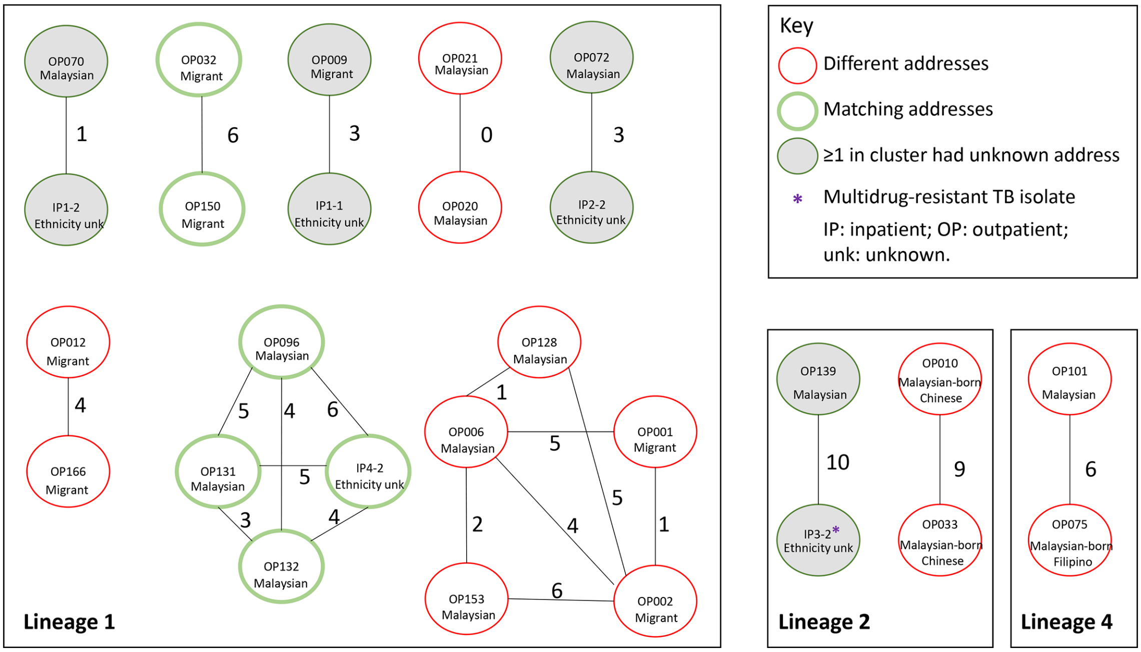

Fig. 5. Plausible transmission links among inpatients and outpatients from Kota Kinabalu, Malaysia, 2012–2017. Genomic clusters were identified using transmission threshold (T)=19 and transmission rate=(β), equivalent to a loose singlenucleotide polymorphism (SNP) threshold of 12. Each circle represents an isolate, while the number linking the circles represents the SNP difference. Ethnicity of individuals from whom isolates were obtained is shown as Malaysian, migrant or unknown. All migrants were Filipino, except the patient who supplied sample OP032 from Brunei.

The catchment area of Luyang outpatient clinic includes impoverished areas with high populations of undocumented migrants. Migrants comprised a third of those involved in TB transmission clusters (Fig. 5). While TB medications are provided free of charge, there are still substantial costs involved in seeking healthcare to determine a TB diagnosis and participate in treatment, so late presentations and undiagnosed and/or unnotified cases occur. Missed opportunities for prevention through contact investigation are also highlighted by the time intervals seen between diagnosis of cases within clusters, noted to span almost the whole study period of nearly 5 years. Since 2012–2017, contact investigation has escalated in Kota Kinabalu, but remains challenging, and is put at risk by intervening public health priorities such as coronavirus disease 2019 (COVID-19).

A finding of note is the greater likelihood of persistence of sputum culture positivity of lineage 2 after 8 weeks of TB treatment compared with lineage 1 strains. Numbers were very small, since not all outpatients could produce a week 8 sample, and for those who could, culture positivity was low overall. However, this finding is supported by previous evidence of lineage 2 being associated with delayed culture conversion [39] and adds further to the knowledge of greater virulence of this TB strain [40], fortunately uncommon in this setting.

Low rates of TB drug resistance were identified in this study overall. Where resistance mutations were identified, high diversity was noted, which illustrates the presence of acquired drug resistance development rather than primary drug resistance in this setting. A novel resistance mutation was identified conferring resistance to isoniazid, which highlights ongoing evolutionary selective pressure on the drug. One isolate had one of the 'disputed' *rpoB* mutations (p.Leu452Pro), which are associated with low-level rifampicin resistance and can be missed by MGIT phenotypic testing [41].

The chief limitation of the study was that the isolates available represented a convenience sample comprising a small proportion of total TB cases diagnosed in the city during the study time frames. Consent was sought to collect demographic data for research purposes from all outpatients, but not from the inpatients whose samples were collected for clinical purposes; therefore demographic data were missing from a proportion. However, the information available for analysis represents the largest dataset from this region to date. There was little temporality in the dataset, probably due to the limited sampling time frame accompanied by a high genomic diversity that could have affected root-to-tip distance, since 179/234 (76%) isolates in the dataset were collected between 2012–2014.

In conclusion, diverse strains of lineage 1 *M. tuberculosis* are seen across patients of all main ethnic groups in this multicultural region of eastern Malaysia. The data have practical public health implications for the Kota Kinabalu setting. Contact investigation strategies should be broadened beyond households, and beyond current restrictions based on age and HIV status, to reduce the likelihood of secondary cases occurring. This would be in keeping with revised World

Health Organization recommendations supporting broader treatment of TB infection [42]. Evidence of several introductions of diverse TB strains highlights the challenge of TB in mobile, cross-border, vulnerable populations. Control in these populations is critical for reducing state-wide and national burdens.

Declarations page

Availability of data and material: Data available on request.

Funding information
This study was funded by grants from the Malaysian Thoracic Society and 'Improving Health Outcomes in the Tropical North: A Multidisciplinary Collaboration (HOT NORTH)', National Health and Medical Research Council (NHMRC) grant number 1131932. Phenotypic susceptibility testing and whole-genome sequencing of TB isolates in sub-study 2 were supported by a grant from National University of Singapore Yong Loo Lin School of Medicine Aspiration Fund (NUHSRO/2014/069/AF-New Idea/04). A. P. R. is supported by NHMRC fellowship 1142011.

Acknowledgements
We thank the Director General of Health, Malaysia, for permission to publish this article. We thank Justine Peh Woei Ling at Singapore General Hospital and Martin Gengerbacher at National University of Singapore for processing samples for sub-studies 1 and 2 respectively. We are grateful to Kelly Nestor, Ann Wee, Thecla Tasius, Florena Maidik and Sharon Nestor for enrolling participants and handling samples; and all the staff at Luyang Outpatient Clinic, particularly Dr Eman Farhad, Dr Maryam Amaleena, Suzie Taylor, Jenny Omar, Dr Elissa Soh and Dr Lee Wai Khew, for supporting the project. We also thank Sushil Pandey at the Queensland Mycobacterium Reference Laboratory for facilitating laboratory work. We also thank the Sabah State Health Department and Clinical Research Centre of Queen Elizabeth Hospital for general assistance.

Conflicts of interest
The authors declare that there are no conflicts of interest.

References

1. World Health Organisation. Global tuberculosis report 2016. who library Cataloguing-in-Publication data. WHO/HTM/TB/2016.13. http://apps.who.int/iris/bitstream/10665/137094/1/9789241564809_eng.pdf?ua=1
2. Dony JF, Ahmad J, Khen Tiong Y. Epidemiology of tuberculosis and leprosy, Sabah, Malaysia. *Tuberculosis* 2004;84:8–18.
3. World Health Organisation. Tuberculosis country profiles. https://www.who.int/tb/country/data/profiles/en/
4. Rashid Ali MRS, Parameswaran U, William T, Bird E, Wilkes CS *et al*. A prospective study of tuberculosis drug susceptibility in Sabah, Malaysia, and an algorithm for management of isoniazid resistance. *J Trop Med* 2015;2015:1–8.
5. William T, Parameswaran U, Lee WK, Yeo TW, Anstey NM *et al*. Pulmonary tuberculosis in outpatients in Sabah, Malaysia: advanced disease but low incidence of HIV co-infection. *BMC Infect Dis* 2015;15:32.
6. Muhammad Redzwan SRA, Ralph AP, Sivaraman Kannan KK, William T. Individualised second line anti-tuberculous therapy for an extensively resistant pulmonary tuberculosis (XDR PTB) in East Malaysia. *Med J Malaysia* 2015;70:200–204.
7. Dale JW, Nor RM, Ramayah S, Tang TH, Zainuddin ZF. Molecular epidemiology of tuberculosis in Malaysia. *J Clin Microbiol* 1999;37:1265–1268.
8. Rashid Ali MRS, Parameswaran U, William T, Bird E, Wilkes CS *et al*. A prospective study of mycobacterial viability in refrigerated, unpreserved sputum batched for up to 8 weeks. *Int J Tuberc Lung Dis* 2015;19:620–621.
9. Ralph AP, Rashid Ali MRS, William T, Piera K, Parameswaran U *et al*. Vitamin D and activated vitamin D in tuberculosis in equatorial Malaysia: a prospective clinical study. *BMC Infect Dis* 2017;17:312.
10. Bainomugisa A, Pandey S, Donnan E, Simpson G, Foster J'Belle *et al*. Cross-Border movement of highly drug-resistant *Mycobacterium tuberculosis* from Papua New Guinea to Australia through Torres Strait protected zone, 2010-2015. *Emerg Infect Dis* 2019;25:406–415.
11. Bainomugisa A, Duarte T, Lavu E, Pandey S, Coulter C *et al*. A complete high-quality MinION nanopore assembly of an extensively drug-resistant Mycobacterium tuberculosis Beijing lineage strain identifies novel variation in repetitive PE/PPE gene regions. *Microb Genom* 2018;4 [Epub ahead of print 15 06 2018].
12. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* 2014;30:2114–2120.
13. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25:1754–1760.
14. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J *et al*. The sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078–2079.
15. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K *et al*. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297–1303.
16. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T *et al*. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* 2012;6:80–92.
17. Merker M, Barbier M, Cox H, Rasigade J-P, Feuerriegel S *et al*. Compensatory evolution drives multidrug-resistant tuberculosis in Central Asia. *elife* 2018;7:e38200 [Epub ahead of print 30 10 2018].
18. Brown TS, Challagundla L, Baugh EH, Omar SV, Mustaev A *et al*. Pre-detection history of extensively drug-resistant tuberculosis in KwaZulu-Natal, South Africa. *Proc Natl Acad Sci U S A* 2019;116:23284–23291.
19. Bainomugisa A, Lavu E, Hiashiri S, Majumdar S, Honjepari A *et al*. Multi-clonal evolution of multi-drug-resistant/extensively drug-resistant Mycobacterium tuberculosis in a high-prevalence setting of Papua New Guinea for over three decades. *Microb Genom* 2018;4.
20. Cirillo DM, Miotto P, Tagliani E, ReSeq TBC, ReSeqTB Consortium. Reaching consensus on drug resistance conferring mutations. *Int J Mycobacteriol* 2016;5 Suppl 1:S33.
21. Coll F, McNerney R, Guerra-Assunção JA, Glynn JR, Perdigão J *et al*. A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. *Nat Commun* 2014;5:4812.
22. Palittapongarnpim P, Ajawatanawong P, Viratyosin W, Smittipat N, Disratthakit A *et al*. Evidence for host-bacterial co-evolution via genome sequence analysis of 480 Thai Mycobacterium tuberculosis lineage 1 isolates. *Sci Rep* 2018;8:11597.
23. Stamatakis A, Hoover P, Rougemont J. A rapid bootstrap algorithm for the RAxML web servers. *Syst Biol* 2008;57:758–771.
24. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the beast 1.7. *Mol Biol Evol* 2012;29:1969–1973.
25. BEAST developers. Using bets to evaluate temporal signal 2021.
26. Duchene S, Lemey P, Stadler T, Ho SYW, Duchene DA *et al*. Bayesian evaluation of temporal signal in measurably evolving populations. *Mol Biol Evol* 2020;37:3363–3379.
27. Walker TM, Ip CLC, Harrell RH, Evans JT, Kapatai G *et al*. Whole-Genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. *Lancet Infect Dis* 2013;13:137–146.
28. Menardo F, Duchêne S, Brites D, Gagneux S. The molecular clock of Mycobacterium tuberculosis. *PLoS Pathog* 2019;15:e1008067.
29. Baele G, Li WLS, Drummond AJ, Suchard MA, Lemey P. Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. *Mol Biol Evol* 2013;30:239–243.

30. Paradis E, Schliep K. Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 2019;35:526–528.
31. Stimson J, Gardy J, Mathema B, Crudu V, Cohen T *et al*. Beyond the SNP threshold: identifying outbreak clusters using inferred transmissions. *Mol Biol Evol* 2019;36:587–603.
32. R Core Team. R: a language and environment for satistical computing. R foundation for statistical computing. https://www.R-project.org/
33. Roetzer A, Diel R, Kohl TA, Rückert C, Nübel U *et al*. Whole genome sequencing versus traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: a longitudinal molecular epidemiological study. *PLoS Med* 2013;10:e1001387.
34. Ford CB, Lin PL, Chase MR, Shah RR, Iartchouk O *et al*. Use of whole genome sequencing to estimate the mutation rate of Mycobacterium tuberculosis during latent infection. *Nat Genet* 2011;43:482–486.
35. Lowbridge C, Fadhil SAM, Krishnan GD, Schimann E, Karuppan RM *et al*. How can gastro-intestinal tuberculosis diagnosis be improved? A prospective cohort study. *BMC Infect Dis* 2020;20:255.
36. Goroh MMD, van den Boogaard CHA, Ibrahim MY, Tha NO, Swe S *et al*. Factors affecting participation of local residents and migrants in tuberculosis contact investigation in a low-income, high-burden setting. *Trop Med Infect Dis* 2020:accepted.
37. Ministry of Health Malaysia. Manual Sistem Maklumat Tibi Kebangsaan (National Tuberculosis Information System [TBIS] Manual) 2002.
38. World Health Organisation. *Treatment of tuberculosis Guidelines Fourth edition WHO/HTM/TB/2009.420*. Geneva, Switzerland: WHO; 2010..
39. Visser ME, Stead MC, Walzl G, Warren R, Schomaker M *et al*. Baseline predictors of sputum culture conversion in pulmonary tuberculosis: importance of cavities, smoking, time to detection and W-Beijing genotype. *PLoS One* 2012;7:e29588.
40. Parwati I, van Crevel R, van Soolingen D. Possible underlying mechanisms for successful emergence of the *Mycobacterium tuberculosis* Beijing genotype strains. *Lancet Infect Dis* 2010;10:103–111.
41. Miotto P, Cabibbe AM, Borroni E, Degano M, Cirillo DM. Role of Disputed Mutations in the *rpoB* Gene in Interpretation of Automated Liquid MGIT Culture Results for Rifampin Susceptibility Testing of Mycobacterium tuberculosis. *J Clin Microbiol* 2018;56 [Epub ahead of print 25 04 2018].
42. WHO. *Latent Tuberculosis Infection: Updated and Consolidated Guidelines for Programmatic Management*. Geneva: WHO Press; 2018.