Lin et al. BMC Infectious Diseases (2020) 20:462
https://doi.org/10.1186/s12879-020-05189-y

BMC Infectious Diseases

RESEARCH ARTICLE Open Access

The geno-spatio analysis of *Mycobacterium tuberculosis* complex in hot and cold spots of Guangxi, China

Dingwen Lin[1†], Zhezhe Cui[1†], Virasakdi Chongsuvivatwong[2*], Prasit Palittapongarnpim[3], Angkana Chaiprasert[4], Wuthiwat Ruangchai[3], Jing Ou[1] and Liwen Huang[1]

Abstract

Background: At present, there are few studies on polymorphism of *Mycobacterium tuberculosis* (Mtb) gene and how it affects the TB epidemic. This study aimed to document the differences of polymorphisms between tuberculosis hot and cold spot areas of Guangxi Zhuang Autonomous Region, China.

Methods: The cold and hot spot areas, each with 3 counties, had been pre-identified by TB incidence for 5 years from the surveillance database. Whole genome sequencing analysis was performed on all sputum Mtb isolates from the detected cases during January and June 2018. Single nucleotide polymorphism (SNP) of each isolate compared to the H37Rv strain were called and used for lineage and sub-lineage identification. Pairwise SNP differences between every pair of isolates were computed. Analyses of Molecular Variance (AMOVA) across counties of the same hot or cold spot area and between the two areas were performed.

Results: As a whole, 59.8% (57.7% sub-lineage 2.2 and 2.1% sub-lineage 2.1) and 39.8% (17.8% sub-lineage 4.4, 6.5% sub-lineage 4.2 and 15.5% sub-lineage 4.5) of the Mtb strains were Lineage 2 and Lineage 4 respectively. The percentages of sub-lineage 2.2 (Beijing family strains) are significantly higher in hot spots. Through the MDS dimension reduction, the genomic population structure in the three hot spot counties is significantly different from those three cold spot counties (T-test $p = 0.05$). The median of SNPs distances among Mtb isolates in cold spots was greater than that in hot spots (897 vs 746, Rank-sum test $p < 0.001$). Three genomic clusters, each with genomic distance ≤12 SNPs, were identified with 2, 3 and 4 consanguineous strains. Two clusters were from hot spots and one was from cold spots.

Conclusion: Narrower genotype diversity in the hot area may indicate higher transmissibility of the Mtb strains in the area compared to those in the cold spot area.

Keywords: Tuberculosis, Genotypes, Polymorphisms, Spatial, Influence

* Correspondence: cvirasak@medicine.psu.ac.th
†Dingwen Lin and Zhezhe Cui contributed equally to this work.
[2]Epidemiology Unit, Faculty of Medicine, Prince of Songkla University, Songkhla 90110, Thailand
Full list of author information is available at the end of the article

© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

It is widely recognized that *Mycobacterium tuberculosis* (Mtb) of different molecular types have different transmission capacities, pathogenicity and drug resistance rates [1, 2]. Therefore, Mtb genotypes may be associated with the tuberculosis (TB) endemicity [3, 4]. Guangxi Zhuang Autonomous Region is a southern region of China with a seriously high TB prevalence [5]. However, the internal TB situation in Guangxi varies greatly. Previous spatiotemporal studies found a significantly high notification spatial cluster (hot spots) and a significantly low notification spatial cluster (cold spots) through spatiotemporal scanning technology [6]. Some environmental and socioeconomic status factors related to the TB epidemics in this region have been identified. However, the molecular biological explanations for this situation are still lacking.

Whole genome sequencing (WGS) was employed in this study because it provided a powerful tool for phylogenetic analysis and epidemiological tracing than other conventional methods such as IS*6110* restriction fragment length polymorphism (RFLP), spoligotyping, and variable-number tandem repeat (VNTR) typing [7–11]. Single nucleotide polymorphism (SNP) based phylogenetic networks of Mtb strains have been used to identify the super-spreaders and transmission [12]. However, the relationship between the number of SNP difference and geographic spread of Mtb has never been studies in depth. Such a study would be feasible in our study since information on Mtb genome of the population and geographic information of the TB case is available.

Based on the above reasons, this study was conducted to 1) document the differences of Mtb polymorphisms (SNP) between hot and cold spot areas, 2) identify Mtb genotype with high level of local transmission and 3) analyse the relationship between different level of SNP variations and geographic distribution of the Mtb isolates.

Methods

Study design

A case-only study in the TB notification hot and cold spot areas of Guangxi was performed from January to June 2018.

Study setting

Based on the spatial clustering analysis, three counties (C1-C3) with significantly high TB notification were identified as hot spots, and three counties (C4-C6) with significantly low notification rate were identified as cold spots [6]. The permanent population (Unit: Thousand) of these counties are 536.2 (C1), 325.6 (C2), 604.2 (C3), 720.7 (C4), 1191.8 (C5) and 795.4 (C6). In these six study sites, patients who were suspected to have pulmonary TB were confirmed at the designated hospitals for TB with chest radiography, sputum smear and culture. All Mtb isolates from the culture were shipped to the Guangxi center for disease prevention and control (GXCDC) where Mtb deoxyribonucleic acid (DNA) extraction was performed before transporting the DNA to Zeta Biosciences company (Shanghai) for WGS and upstream data analysis. After receiving informed consents, epidemiological investigations of the TB patients were conducted by local hospital teams under the supervision of GXCDC. Clinical and other laboratory data of the recruited patients were retrieved from routinely entered data of the National Notifiable Disease Reported System which was overseen by GXCDC.

Study participants and selection methods

Sample size

The sample size of this study was calculated based on the formula of two independent proportions comparison [13]. According to a previous study of dominant genotype (Beijing family strains) in Guangxi, the estimated percentage of Beijing strains was 70% in hot spot areas and 50% in cold spot areas. With a type I error of 0.05 and a power of 90%, at least 248 active TB cases (124 cases in each spots) were required.

Eligibility criteria for index cases

Eligible index TB cases must have been a resident in the study sites for at least 2 years prior to TB diagnosis. Isolates from individuals who were unable to communicate with investigators and children under the age of five were excluded.

WGS performance and SNP calling

At the GXCDC, Mtb DNA extraction was conducted by a genetic sample kit (HiPure Bacterial DNA Kit, Magen Biotech Co. Ltd). At Zeta Biosciences (Shanghai), a WGS kit was employed to obtain enough nucleic acid for the sequencing of the downstream analysis. Each sample was quantified by Qubit 2.0 Fluorometer (Invitrogen, Carlsbad, CA, USA). Next generation sequencing library with 350-base-pair (bp) paired-end preparations were constructed for each purified DNA sample according to the manufacturer's criterion (Illumina TruSeq DNA Nano Library Prep Kit). Then libraries with different indices were multiplexed and loaded on an Illumina HiSeq instrument with an expected coverage of 100 following the manufacturer's instructions (Illumina, San Diego, CA, USA). After analyzing a short sequence alignment of the Mtb according to the reference genome H37Rv, multiple sequence alignment test strains of SNPs/InDels were used to obtain the corresponding molecular classification [14]. A Burrows-Wheeler transform algorithm and genome analysis toolkit packages (GATK v

4.1.1.0, Broad Institute, USA) were applied in this process. The SNVs that were present in any drug-resistance gene, mobile genetic element, phage, PE/PPE region and non-homozygous SNVs were discarded. The remaining SNVs had been converted to an SNV-supermatrix using an in-house Python script before being used in the phylogenetic analysis. Phylogenetic trees were constructed by Bayesian Inference (BI) methods using MrBayes [15]. The BI tree was supported with posterior probabilities. The tree was visualized by FigTree version 1.4.2. Molecular typing and statistical inference were conducted based on the genotype assignment of the isolates were based on SNPs classification as previously published [16] Finally, the crude and filtered fasta-files, vcf files, SNPs distance matrix, classified spoligotype and lineage genotype data, and the Weir and Cockerham weighted genetic group structure differentiation coefficient (F-statistics, Fst) between the subgroups were delivered to the GXCDC for downstream data analysis.

Data management and analysis

Epi Data (v 3.1) was used for double entering the data at regular intervals while R (v 3.3.2) was used for epidemiological data management and analysis. MEGA-X (v 10.0.5) and Fig Tree (v 1.4.4) were used to build phylogenetic trees with color labels. Multidimensional scaling model (MDS) was employed to test the between and within group of differentiation *Fst* value from Analysis of Molecular Variance (AMOVA) by reducing the matrix dimensions. Clustering of isolates were analyzed based on several levels of SNP distances starting from 12 as generally done but were also done at the SNP distances of 24, 48 and 96, and so on. Isolates with less than 12 SNP distances were considered as recent epidemiologically linked. Clusters of isolates with higher distances were analyzed in regard to geographical proximity. Cluster of isolates having SNP difference within each of these cut points were identified and checked to see whether all clusters members were from the same county, or area (Hot and cold spots). Rank-sum test, a non-parametric analysis, was employed to compare the median of pairwise SNPs distance between two groups.

Results

After excluding 13 participants (8 patients with non-tuberculosis mycobacteria infection and 5 patients with low quality of Mtb DNA), a total of 147 isolates from hot spots and 144 from cold spots were included for further analysis. Their phylogenetic trees are shown in Fig. 1. The predominate lineage is lineage 2 (59.8%), with its major sub-lineage 2.2 or the Beijing family (57.7%). The percentage of sub-lineage 2.1 (2.1%) was relatively high compared to other part of China as previously reported [17]. Other major genotype included lineage 4 (39.8%) with its sub-lineage 4.4 (17.8%), 4.2 (6.5%) and 4.5 (15.5%). Three most likely clusters were detected at the criteria of SNPs distance less than or equal to 12.

Table 1 shows the lineage distribution in each county while Table 2 compared the Beijing family with all other strains. The proportion of lineage 2.2.1 and 2.2.2 (Beijing family strains) was evenly distributed in the three hot counties, and was significantly higher than that in the cold counties ($p = 0.022$). Compared with the hot spots, the isolates from the cold spots are more genetically diverse. There was no significant difference of Beijing family percentages between the counties in each spot ($p = 0.361$ and 0.482).

As shown in Fig. 2, after AMOVA computing and filtering, 14,250 SNP sites were kept for Fst estimation between hot and cold spots. The average Fst value with Weir and Cockerham weighted is 0.0195. The SNP sites which have the highest Fst estimation are within the coding sequences of Rv1186c (PruC) (0.133), and hypothetical proteins, Rv0210 (0.101), Rv1508c (0.1074) and Rv3900c (0.112). PruC is a membrane-associated DNA-binding protein that control proline metabolism.

The Multidimensional Scaling analysis was used to compare the Fst of tuberculosis strains across six counties as shown in Fig. 3. After introducing the Fst matrix, we performed the MDS and reduced the distance dimension to one to see the similar scores between these counties, and mapped it. The Fst distance similarity scores have certain clustering property according to different spatial spots. The mean similarity scores of three hot spot counties is 0.013 ± 0.012, while the mean score for the three cold spots was -0.013 ± 0.008. The difference between two groups is significant ($p = 0.05$).

Among 42,195 pairs of SNPs difference, the shortest distance between these stains is 7 SNPs, and the longest distance pair was 1893 SNPs. The average distance was 703 SNPs. Figure 4 shows the frequency of SNPs distance (one to one comparison) in hot and cold spot areas. Both cold and hot spots had three peaks of SNPs distances in the range of 0–500, 501–1000 and > 1000. The leftmost peaks included the distances of recently transmitted isolates. The median of SNPs distance among isolates from cold spots is greater than that in hot spot areas (897 vs 746, Rank-sum test $p < 0.001$). This is consistent with the higher transmission activities in the hot spot.

Table 3 shows the relationship of SNPs based clusters within and across the hot and cold spots at different SNPs level. Two isolates having SNPs difference of less than or equal to 12 were likely to be related by recent transmission. We detected two genetic clusters with 2 and 3 consanguineous strains in a hot spot county and one genetic cluster of 4 consanguineous strains cross two cold spot counties (Supplementary Table 1.). Only if the clustering criterion was relaxed to 96 of SNPs distance, we would

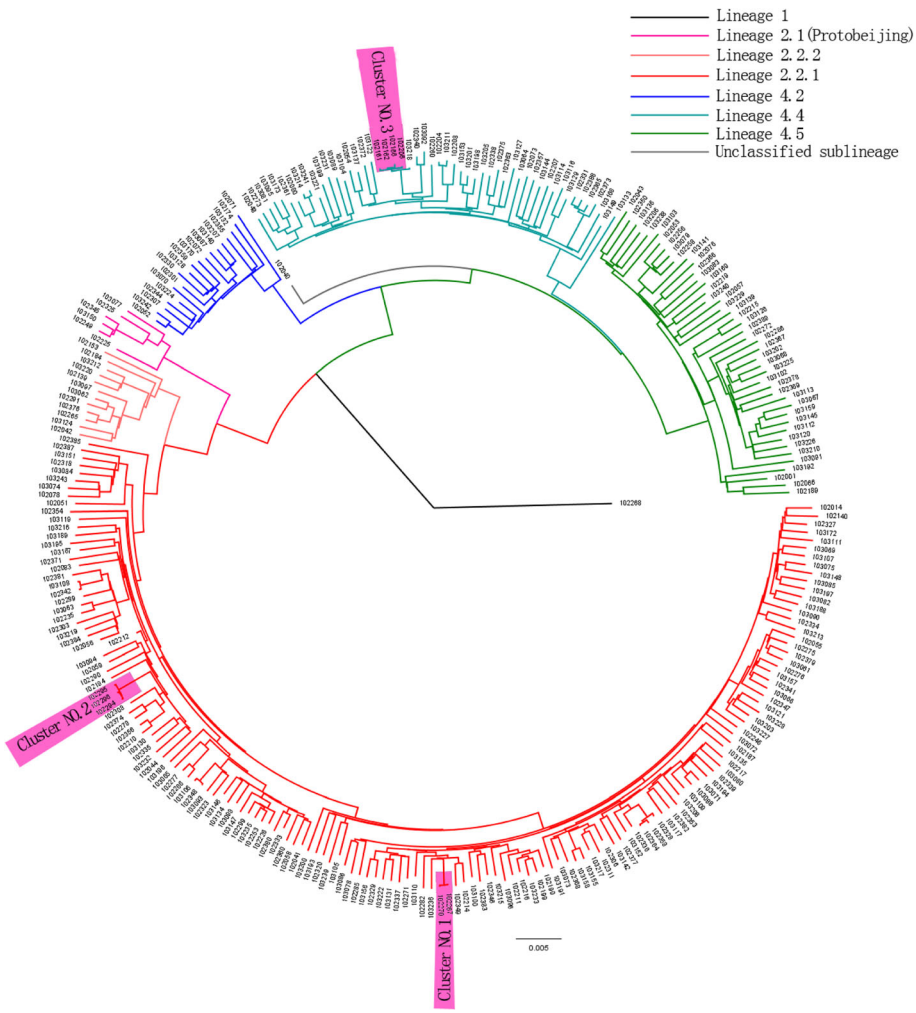

Fig. 1 The phylogenetic tree of Mtb constructed by the Bayesian Inference method. The three most likely recent clusters,with the criterion of SNP distances less than or equal to 12, were shaded. The clusters No. 1 & 2 were found in a county of hot spots and the cluster No. 3 was found in two counties of cold spots. Only a single Lineage 1 isolate (102268) was identified

identify the 3 more consanguineous strains clusters across two zones. The number of clusters, geometric mean of cluster size, number of cluster crossing the county and number of cluster crossing the zone have positive correlations with cut point level ($p < 0.05$).

Discussion

This study identified the molecular biological characteristics of Mtb in different TB epidemic areas of southern China. Only two major lineages (L2 and L4) were found in this study setting. The predominate Mtb strain is

Table 1 Lineage distribution at county level(n, row%)

Region	Subgroups	lineage1.1.1.1	lineage2.1	lineage2.2.1		lineage2.2.2	lineage4.2.2	lineage4.4.1	lineage4.4.2	lineage4.5	Total
				Ancestral	Modern						
Hot spots	C1	0 (0)	1 (1.5)	14 (21.54)	27 (41.54)	2 (3.1)	2 (3.1)	0 (0)	8 (12.3)	11 (16.9)	65 (100)
	C2	0 (0)	1 (2.4)	9 (21.95)	13 (31.71)	1 (2.4)	4 (9.8)	0 (0)	10 (24.4)	3 (7.3)	41 (100)
	C3	1 (2.4)	3 (7.3)	17 (41.46)	10 (24.39)	2 (4.9)	2 (4.9)	0 (0)	1 (2.4)	5 (12.2)	41 (100)
Cold spots	C4	0 (0)	0 (0)	12 (20.34)	18 (30.51)	3 (5.1)	3 (5.1)	0 (0)	12 (20.3)	11 (18.6)	59 (100)
	C5	0 (0)	1 (1.7)	9 (15.52)	15 (25.86)	2 (3.4)	5 (8.6)	1 (1.7)	14 (24.1)	11 (19)	58 (100)
	C6	0 (0)	0 (0)	9 (33.33)	3 (11.11)	2 (7.4)	3 (11.1)	0 (0)	6 (22.2)	4 (14.8)	27 (100)
Total		1 (0.3)	6 (2.1)	70 (24.05)	86 (29.55)	12 (4.1)	19 (6.5)	1 (0.3)	51 (17.5)	45 (15.5)	291 (100)

Table 2 Comparison the proportion of Beijing genotype in each group

Region	Subgroups	Beijing	Non-Beijing	*p* value (Compare counties in each spots)	*p* value (Compare hot and cold spots)
Hot spots	C1	43	22	0.361	0.022
	C2	23	18		
	C3	29	12		
Cold spots	C4	33	26	0.482	
	C5	26	32		
	C6	14	13		

lineage 2.2 (Beijing family), and it was significantly higher in TB notification hot spots. Through the population gene structure analysis (AMOVA) and SNPs comparison between cold and hot spots and the multidimensional scaling modeling of each county, we found that the two spot areas had some differences in genetic structure, and the spatial internal consistency was relatively high. Specific SNPs sites between the cold and hot spots with high Fst estimation mapped to special proteins that may contribute to the pathogenicity differences in Mtb. Three genomic (SNPs ≤12) and geographic groups were detected and identified as Mtb recent transmission individuals.

Previous studies have suggested that *Mtb* among human originated in Africa and was divided into seven lineages by several thousand years of mutations [18–21]. The evolution of Mtb has been related to human migration and evolution. It spread from Africa to the rest of the world along with human migration and formed the current genotype distribution. Nowadays, the most prevalent Mtb strain in China is lineage 2 [22]. Although in northern China, the proportion of Beijing strains is as high as 80%, which were mostly Modern Beijing strains), as the latitude decreases, this proportion decreases [23]. Moreover, with the increase of population mobility, the polymorphism of *Mtb* genotype becomes more and more obvious. Therefore, as a southern province of China, Guangxi has more proportion of Ancestral Beijing strains and genetic diversity of Mtb strains than that in northern region [24]. The origin of Protobeijing strain (L 2.1) is likely to be in Southern China as it has the highest percentage [16]. In this study, the lineage 4 with three sub-lineages (L4.2, L4.4 and L4.5) also accounted for a large proportion. In contrast to other major human-adapted lineages, lineage 4 appears with significant frequency on all inhabited continents [25]. Thus, it is the most widespread cause of TB in humans geographically [26]. Among this lineage, L4.4 and L4.5 were mostly reported from China, although we usually called L4 as Euro-American [27]. Stucki hypothesized that the global spread of L4 maybe caused by European migration and colonization [27]. Yet, the reasons for this spatial distribution in China needs more evidence.

In this study, by comparing all the gene loci of the cold and hot spots strain, it was found that the two populations had mutation differences in some special regulatory proteins. The mutation of Rv1186c (PruC) has a certain significance. Mtb is an obligate aerobic bacterium that needs oxygen to grow. However, paradoxically, it shows a remarkable metabolic flexibility that allows

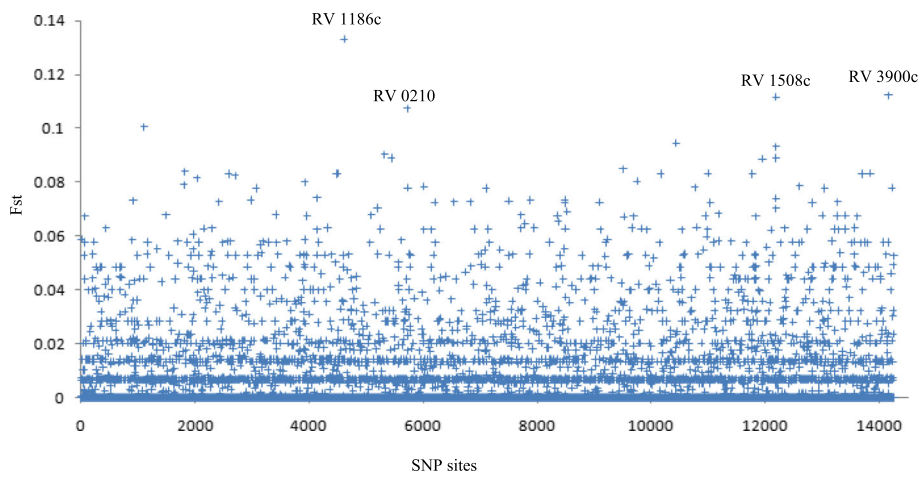

Fig. 2 The Fst of each SNP sites of MTB isolates in hot spots compared with cold spots (Weir and Cockerham weighted). The SNP sites that showed highest Fst are labeled

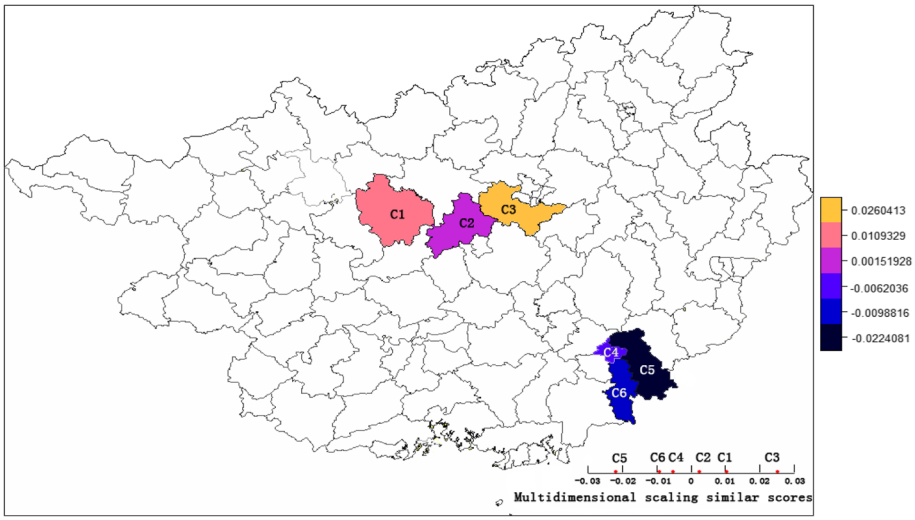

Fig. 3 The Multidemensional scaling similar scores of Fst between six counties. (We are very grateful to Center for Spatial Sciences at the University of California, Davis for providing us with the map files which are freely available for academic use. https://www.gadm.org/about.html)

itself to survive and metabolize in oxygen-deprived conditions [28]. It has been shown that mycobacteria can grow on proline as the sole carbon and energy source under hypoxia, and it is regulated by a unique transcriptional regulator (PruC) [29]. An animal study performed by Smith DA et al. found that mycobacteria with abnormal proline metabolism were nonpathogenic in immune-competent mice [30]. However, Rv1186c was predicted to be non-essential gene in the papers of DeJesus et al. [31] and Lamichhane et al. [32], except the paper of Sassetti et al. [33]. Apart from Rv1186c, the 3 genes (Rv0210, Rv1508c, Rv3900c) that came out of the SNP analysis are unknown function genes and all are predicted as non-essential genes. Thus, it is expected that there will be no significant association with the pathogenicity of Mtb. Nevertheless, further epidemiological and clinical exploration are needed.

As the dominant genotype, Beijing family strains have been shown to cluster more frequently [34]. This suggests that Mtb recent transmission is more likely to occur in such strains [35, 36]. Some scholars claimed that the determination of recent transmission or MTB outbreak (transmission within 2 or 3 years) is that the cut-offs of WGS-Based genomic distance is less than or equal to 12 SNPs [37]. However, only three recent transmission groups in this study were detected. In the previous research, the research samples with recent transmission cases are generally from communities of long-term surveillance or tuberculosis outbreaks field [12, 38]. Although the specimens in this study were from two spatial clusters

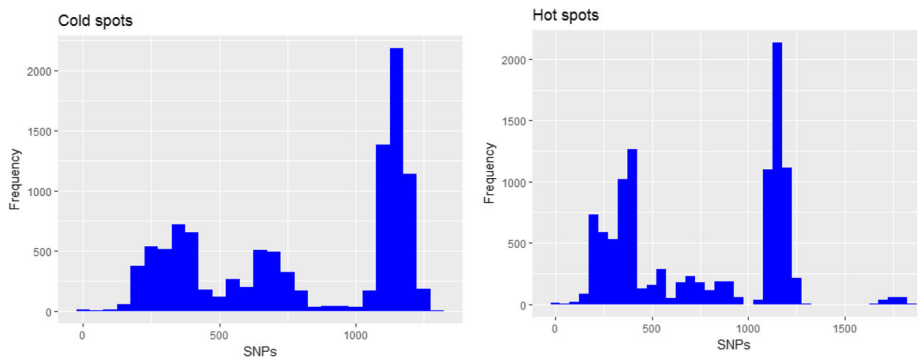

Fig. 4 The frequency distribution of SNPs distances in hot and cold spot areas. In general, SNP distance peak of > 1000 are form differences between isolates belonging to different lineages, in this case between L2 and L4. 500–1000 are resulted from major sub-lineages such as between L2.1 and L2.2. The ones < 500 are usually from differences between more detailed sub-lineages such Asian African2 and Asian African 3. < 100 is almost definitely from the difference between isolates belonging to the same detailed sub-lineages or the same cluster

Table 3 SNPs based clustering at different clustering criterias

SNP difference cut point	≤ 12	≤ 24	≤ 48	≤ 96	≤ 192	≤ 384
No. of clusters	3	6	8	20	16	27
Geometric mean (geometric SD) of cluster size	2.88 (0.82)	1.92 (1.17)	2.48 (1.06)	2.38 (1.23)	5.87 (15.90)	36.77 (57.12)
No. of cluster crossing counties in the same zone	1	4	4	10	13	24
No. of cluster acrossing zones	0	0	0	3	13	24

of TB notification (hot and cold spots), it is likely to be true that there is no obvious outbreak occurring during the study period. The locations of included participants were scattered. Thus, this study showed that the median SNPs distance of strains in hot spots was significantly lower than that in cold spots. The comparison of SNPs population genetic structure was also proved the significant difference in the gene structure between the two areas, but the differences within the areas were relatively small. We did not find any cluster that have members crossing the hot and cold spots. Actually the minimal SNP distance between any genetically related isolates in both spots were at least 96. This suggests that the transmission pattern of the Mtb in hot spots may be different from those in cold spots. Local transmission in hot spot areas (over a period of more than 3 years) is more likely than in cold spot areas. Homologous transmission may occur over a longer period of time [37]. They might have gotten the same type of *Mycobacterium tuberculosis* many years ago, and the strains might have mutated after a long time of latent infection, proliferation and then endogenous reactivation. Thus, the SNPs distance between the two strains would become larger. Our study also performed contact tracing among included cases [39]. The results showed that the detection rate of TB in household contacts was very low. Only two domestic cases have been detected in hot spot areas. Therefore, we believe that the transmission patterns of TB patients in cold and hot spots were dominated by community transmission.

Either the spreading of the Mtb was local or there were some socioeconomic factors that hinder the transmission between the hot and cold spot areas. Our previous studies on the ecology of tuberculosis suggested that there is a negative correlation between average sunshine time and reported incidence of tuberculosis [6]. Therefore, there may be some interaction between natural factors and strain pathogenicity. This requires further exploration.

Meanwhile, sub-lineage 2.1 (Protobeijing strain), a special subgroup related to drug resistance, is mainly concentrated in hot spots. Since it has been reported that the virulence and drug resistance of Beijing gene strain is greater than other strains [34, 40, 41], we can infer that the prevalence of this strain in hot spots is one of the important reasons for its higher TB epidemic than cold spots.

The limitation of this study is that the sputum isolates were collected mainly from passively detected cases in the public hospital. This may have contributed to the fact that some specimens of the TB cases treated in private practice were not included during the study period, which caused the bias to some extent. However, as TB case management strategies become more widely publicized under the national TB control programme, this impact should diminish [42, 43]. The estimated inclusion rate were over 90%.

Conclusions

Mtb genotype distribution is associated with TB incidence. Hot spot area in Guangxi is associated with Beijing family predomination. Narrower genotype diversity in the hot area may indicate higher transmissibility of the Mtb strains in the area compared to those in the cold spot area. These findings demonstrate that promotion of the genetic diagnosis in tuberculosis clinics and early identification of Beijing family strain should be considered in tracking and stopping TB transmission.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12879-020-05189-y.

Additional file 1: Supplementary Table 1. Genetic clustering based on the SNPs (≤12)

Abbreviations

AMOVA: Analyses of Molecular Variance; BI: Bayesian Inference; CDC: Center for DiseaseControl and Prevention; DNA: deoxyribonucleic acid; Fst: F-statistics; MDS: Multidimensional scaling model; Mtb: *Mycobacterium tuberculosis*; RFLP: Restriction Fragment Length Polymorphism; SNP: Single Nucleotide Polymorphism; TB: Tuberculosis; VNTR: Variable-Number Tandem Repeat; WGS: Whole genome sequencing.

Acknowledgments

We would like to thank all the health workers for their assistance in performing the survey. English grammar was revised by Rejean Ruiel Regis Sawh from Health Science Center, Peking University, Beijing, China.

Authors' contributions

VC, DL and ZC contributed to the study design, preparation of materials, data collection, data analysis, and manuscript writing. PP contributed to the study design and manuscript writing. AC and WR contributed to the manuscript writing. JO and LH contributed to the data collection. The author(s) read and approved the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (81760603), Guangxi Natural Science Foundation

(2018GXNSFAA281018) and Guangxi health promotion project (S2019067). The study was part of the Ph.D. thesis of the second author to fulfill the requirement for the tuberculosis research training program at the Epidemiology Unit, Prince of Songkla University, under the support of the Fogarty International Center, National Institutes of Health (D43TW009522).

Availability of data and materials
The raw sequence data reported in this paper have been deposited in the Genome Sequence Archive (Genomics, Proteomics & Bioinformatics 2017) in BIG Data Center (Nucleic Acids Res 2019), Beijing Institute of Genomics (BIG), Chinese Academy of Sciences, under accession numbers PRJCA002021 that are publicly accessible at https://bigd.big.ac.cn/gsa.

Ethics approval and consent to participate
The study protocol was approved by the Institutional Review Board of Guangxi Center for Disease Control and Prevention (GXCDC) (GW-2017-0001) and the Research Ethics Committee of Prince of Songkla University (60–286–18-6). Written informed consents to participate have been provided. All the records were anonymized.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Tuberculosis Control, Guangxi Zhuang Autonomous Region Center for Disease Control and Prevention, Nanning 530028, Guangxi, China. [2]Epidemiology Unit, Faculty of Medicine, Prince of Songkla University, Songkhla 90110, Thailand. [3]Pornchai Matangkasombut Center of Microbial Genomics, Department of Microbiology, Faculty of Science, Mahidol University, Bangkok 10700, Thailand. [4]Office for Research and Development, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok 10700, Thailand.

Received: 11 February 2020 Accepted: 22 June 2020
Published online: 01 July 2020

References
1. Maharjan B, Nakajima C, Isoda N, Thapa J, Poudel A, Shah Y, et al. Genetic diversity and distribution dynamics of multidrug-resistant mycobacterium tuberculosis isolates in Nepal. Sci Rep-Uk. 2018;8:16634.
2. Wiens KE, Woyczynski LP, Ledesma JR, Ross JM, Zenteno-Cuevas R, Goodridge A, et al. Global variation in bacterial strains that cause tuberculosis disease: a systematic review and meta-analysis. BMC Med. 2018;16:196.
3. Monteserin J, Paul R, Gravina E, Reniero A, Hernandez T, Mazzeo E, et al. Genotypic diversity of mycobacterium tuberculosis in Buenos Aires, Argentina. Infection Genetics Evol. 2018;62:1–7.
4. Rasigade JP, Barbier M, Dumitrescu O, Pichat C, Carret G, Ronnaux-Baron AS, et al. Strain-specific estimation of epidemic success provides insights into the transmission dynamics of tuberculosis. Sci Rep-Uk. 2017;7:45326.
5. Lin M, Cui ZZ, Lin DW, Liang DB, Huang MY, Su HB, et al. Visual-spatial and temporal characteristics related to infectious tuberculosis epidemics in Guangxi Zhuang autonomous region, 2012-2015. Zhonghua Liu Xing Bing Xue Za Zhi. 2017;38:1206–11.
6. Cui Z, Lin D, Chongsuvivatwong V, Zhao J, Lin M, Ou J, et al. Spatiotemporal patterns and ecological factors of tuberculosis notification: a spatial panel data analysis in Guangxi. China PLoS One. 2019;14:e0212051.
7. Gardy JL, Johnston JC, Sui SJH, Cook VJ, Shah LN, Brodkin E, et al. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. New Engl J Med. 2011;364:730–9.
8. Hanekom M, van der Spuy GD, Gey van Pittius NC, McEvoy CRE, Hoek KGP, Ndabambi SL, et al. Discordance between mycobacterial interspersed repetitive-unit-variable-number tandem-repeat typing and IS6110 restriction fragment length polymorphism genotyping for analysis of mycobacterium tuberculosis Beijing strains in a setting of high incidence of tuberculosis. J Clin Microbiol. 2008;46:3338–45.
9. Luo T, Yang CG, Gagneux S, Gicquel B, Mei J, Gao Q. Combination of single nucleotide polymorphism and variable-number tandem repeats for genotyping a homogenous population of mycobacterium tuberculosis Beijing strains in China. J Clin Microbiol. 2012;50:633–9.
10. Mathema B, Kurepina NE, Bifani PJ, Kreiswirth BN. Molecular epidemiology of tuberculosis: current insights. Clin Microbiol Rev. 2006;19:658–85.
11. Roetzer A, Diel R, Kohl TA, Ruckert C, Nubel U, Blom J, 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.
12. Walker TM, Clp CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, et al. Whole-genome sequencing to delineate mycobacterium tuberculosis outbreaks: a retrospective observational study. Lancet Infect Dis. 2013;13:137–46.
13. Fleiss JL, Levin B, Paik MC. Statistical methods for rates and proportions. 3rd ed. Hoboken, N.J.: J. Wiley, 2003 Wiley series in probability and statistics).
14. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10:R25.
15. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 2001;17:754–5.
16. Ajawatanawong P, Yanai H, Smittipat N, Disratthakit A, Yamada N, Miyahara R, et al. A novel ancestral Beijing sublineage of mycobacterium tuberculosis suggests the transition site to modern Beijing sublineages. Sci Rep. 2019;9:13718.
17. Luo T, Comas I, Luo D, Lu B, Wu J, Wei L, et al. Southern east Asian origin and coexpansion of mycobacterium tuberculosis Beijing family with Han Chinese. Proc Natl Acad Sci U S A. 2015;112:8136–41.
18. Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, et al. Out-of-Africa migration and Neolithic coexpansion of mycobacterium tuberculosis with modern humans. Nat Genet. 2013;45:1176–82.
19. Gutierrez MC, Brisse S, Brosch R, Fabre M, Omais B, Marmiesse M, et al. Ancient origin and gene mosaicism of the progenitor of mycobacterium tuberculosis. PLoS Pathog. 2005;1:e5.
20. Hershberg R, Lipatov M, Small PM, Sheffer H, Niemann S, Homolka S, et al. High functional diversity in mycobacterium tuberculosis driven by genetic drift and human demography. PLoS Biol. 2008;6:e311.
21. Wirth T, Hildebrand F, Allix-Beguec C, Wolbeling F, Kubica T, Kremer K, et al. Origin, spread and demography of the mycobacterium tuberculosis complex. PLoS Pathog. 2008;4:e1000160.
22. Liu Q, Ma A, Wei L, Pang Y, Wu B, Luo T, et al. China's tuberculosis epidemic stems from historical expansion of four strains of mycobacterium tuberculosis. Nat Ecol Evol. 2018;2:1982–92.
23. Guo YL, Liu Y, Wang SM, Li CY. The identification of mycobacterium tuberculosis isolates by DNA typing technique. Zhonghua Liu Xing Bing Xue Za Zhi. 2005;26:361–5.
24. Luo D, Zhao J, Lin M, Liu F, Huang S, Zhang Y, et al. Drug resistance in newly presenting and previously treated tuberculosis patients in Guangxi Province, People's Republic of China. Asia Pac J Public Health. 2017;29:296–303.
25. Demay C, Liens B, Burguiere T, Hill V, Couvin D, Millet J, et al. SITVITWEB--a publicly available international multimarker database for studying mycobacterium tuberculosis genetic diversity and molecular epidemiology. Infect Genet Evol. 2012;12:755–66.
26. Coscolla M, Gagneux S. Consequences of genomic diversity in mycobacterium tuberculosis. Semin Immunol. 2014;26:431–44.
27. Stucki D, Brites D, Jeljeli L, Coscolla M, Liu Q, Trauner A, et al. Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages. Nat Genet. 2016;48:1535–43.
28. Berney M, Cook GM. Unique flexibility in energy metabolism allows mycobacteria to combat starvation and hypoxia. PLoS One. 2010;5:e8614.
29. Berney M, Weimar MR, Heikal A, Cook GM. Regulation of proline metabolism in mycobacteria and its role in carbon metabolism under hypoxia. Mol Microbiol. 2012;84:664–81.
30. Smith DA, Parish T, Stoker NG, Bancroft GJ. Characterization of auxotrophic mutants of mycobacterium tuberculosis and their potential as vaccine candidates. Infect Immun. 2001;69:1142–50.
31. DeJesus MA, Gerrick ER, Xu W, Park SW, Long JE, Boutte CC, et al. Comprehensive essentiality analysis of the mycobacterium tuberculosis genome via saturating transposon mutagenesis. Mbio. 2017;8:e02133–16.
32. Lamichhane G, Zignol M, Blades NJ, Geiman DE, Dougherty A, Grosset J, et al. A postgenomic method for predicting essential genes at subsaturation levels of mutagenesis: application to mycobacterium tuberculosis. Proc Natl Acad Sci U S A. 2003;100:7213–8.

33. Sassetti CM, Rubin EJ. Genetic requirements for mycobacterial survival during infection. Proc Natl Acad Sci U S A. 2003;100:12989–94.
34. Liu Y, Zhang X, Zhang Y, Sun Y, Yao C, Wang W, et al. Characterization of mycobacterium tuberculosis strains in Beijing, China: drug susceptibility phenotypes and Beijing genotype family transmission. BMC Infect Dis. 2018; 18:658.
35. Wang W, Hu Y, Mathema B, Jiang W, Kreiswirth B, Xu B. Recent transmission of W-Beijing family mycobacterium tuberculosis in rural eastern China. Int J Tuberc Lung Dis. 2012;16:306–11.
36. Yamamoto K, Takeuchi S, Seto J, Shimouchi A, Komukai J, Hase A, et al. Longitudinal genotyping surveillance of *Mycobacterium tuberculosis* in an area with high tuberculosis incidence shows high transmission rate of the modern Beijing subfamily in Japan. Infect Genet Evol 2018;25–30.
37. Meehan CJ, Moris P, Kohl TA, Pecerska J, Akter S, Merker M, et al. The relationship between transmission time and clustering methods in mycobacterium tuberculosis epidemiology. Ebiomedicine. 2018;37:410–6.
38. Luo T, Yang C, Peng Y, Lu L, Sun G, Wu J, et al. Whole-genome sequencing to detect recent transmission of mycobacterium tuberculosis in settings with a high burden of tuberculosis. Tuberculosis (Edinb). 2014;94:434–40.
39. Cui Z, Lin D, Chongsuvivatwong V, Graviss EA, Chaiprasert A, Palittapongarnpim P, et al. Hot and cold spot areas of household tuberculosis transmission in southern China: effects of socio-economic status and mycobacterium tuberculosis genotypes. Int J Environ Res Public Health. 2019;16:1863.
40. Hoffner S, Sahebi L, Ansarin K, Sabour S, Mohajeri P. Mycobacterium tuberculosis of the Beijing genotype in Iran and the World Health Organization eastern Mediterranean region: a meta-analysis. Microb Drug Resist. 2018;24:693–8.
41. Reiling N, Homolka S, Walter K, Brandenburg J, Niwinski L, Ernst M, et al. Clade-specific virulence patterns of mycobacterium tuberculosis complex strains in human primary macrophages and Aerogenically infected mice. Mbio. 2013;4:e00250–13.
42. Li J, Liu XQ, Jiang SW, Li X, Yu F, Wang Y, et al. Improving tuberculosis case detection in underdeveloped multi-ethnic regions with high disease burden: a case study of integrated control program in China. Infect Dis Poverty. 2017;6:151.
43. Wang LX, Liu XQ, Huang F, Hennig C, Uplekar M, Jiang SW. Engaging hospitals to meet tuberculosis control targets in China: using the internet as a tool to put policy into practice. B World Health Organ. 2010;88:937–42.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC