Tuberculosis 125 (2020) 102010

Contents lists available at ScienceDirect

Tuberculosis

journal homepage: http://www.elsevier.com/locate/tube

Investigating extradomiciliary transmission of tuberculosis: An exploratory approach using social network patterns of TB cases and controls and the genotyping of *Mycobacterium tuberculosis*

Suani T.R. Pinho [a], Susan M. Pereira [b], José G.V. Miranda [a], Tonya A. Duarte [c,*], Joilda S. Nery [b], Maeli G. de Oliveira [d], M. Yana G.S. Freitas [d], Naila A. De Almeida [e], Fabio B. Moreira [a], Raoni B. C. Gomes [b], Ligia Kerr [f], Carl Kendall [g], M. Gabriela M. Gomes [h], Theolis C.B. Bessa [i], Roberto F. S. Andrade [a], Mauricio L. Barreto [j]

[a] *Instituto De Física – UFBA. R. Barão De Jeremoabo, S/n. Ondina, 40170-115, Salvador, BA, Brazil*
[b] *Instituto De Saúde Coletiva – UFBA. R. Basílio da Gama, S/n. Canela, 40110-040, Salvador, BA, Brazil*
[c] *Instituto De Ciências da Saúde – UFBA. Av. Reitor Miguel Calmon, S/n. Canela, 40231-300, Salvador, BA, Brazil*
[d] *Universidade Estadual De Feira De Santana. Av. Transnordestina, S/n. Novo Horizonte, 44036-900, Feira de Santana, BA, Brazil*
[e] *Serviço Nacional De Aprendizagem Industrial – SENAI. R, Henrique Dias. Roma, 40444-000, Salvador, BA, Brazil*
[f] *Faculdade De Medicina – UFC. R. Alexandre Baraúna, 949. Rodolfo Teófilo, 60430-160, Fortaleza, CE, Brazil*
[g] *School of Public Health and Tropical Medicine Tulane University, 1440 Canal St, New Orleans, LA, 70112, United States*
[h] *Liverpool School of Tropical Medicine, Liverpool, UK, Pembroke Pl, Liverpool L3 5QA, Reino Unido, UK*
[i] *Instituto Gonçalo Moniz – IGM/FIOCRUZ. R. Waldemar Falcão, 121. Candeal, 40296-710, Salvador, BA, Brazil*
[j] *Centro de Integração de Dados e Conhecimentos para Saúde – CIDACS/FIOCRUZ, Parque Tecnológico Edf. Tecnocentro. Rua Mundo, 121, Salvador, BA, Brazil*

ARTICLE INFO

Keywords:
Tuberculosis
Infection
Social networks
Molecular epidemiology

ABSTRACT

Extradomiciliary contacts have been overlooked in the study of TB transmission due to difficulties in identifying actual contacts in large populations. Complex network analysis provides a framework to model the structure of contacts, specially extradomiciliary ones. We conducted a study of incident sputum-positive TB cases and healthy controls occurring in a moderate TB burden city. Cases and controls were interviewed to obtain data regarding the usual locations of residence, work, study, and leisure. *Mycobacterium tuberculosis* isolated from sputum was genotyped. The collected data were used to build networks based on a framework of putative social interactions indicating possible TB transmission. A user-friendly open source environment (GraphTube) was setup to extract information from the collected data. Networks based on the likelihood of patient-patient, patient-healthy, and healthy-healthy contacts were setup, depending on a constraint of geographical distance of places attended by the volunteers. Using a threshold for the geographical distance of 300 m, the differences between TB cases and controls are revealed. Several clusters formed by social network nodes with high genotypic similarity were characterized. The developed framework provided consistent results and can be used to support the targeted search of potentially infected individuals and to help to understand the TB transmission.

(*continued on next column*)

(*continued*)

(*continued on next page*)

* Corresponding author.
E-mail addresses: suani@ufba.br (S.T.R. Pinho), susanmp@ufba.br (S.M. Pereira), vivas@ufba.br (J.G.V. Miranda), tonya.duarte@gmail.com (T.A. Duarte), joildanery@gmail.com (J.S. Nery), maelioli@hotmail.com (M.G. de Oliveira), yana.guimaraess@gmail.com (M.Y.G.S. Freitas), nailalves@gmail.com (N.A. De Almeida), fabiobm1313@gmail.com (F.B. Moreira), raonidonordeste@gmail.com (R.B.C. Gomes), ligia@ufc.br (L. Kerr), ckendall@tulane.edu (C. Kendall), gabriela.gomes@lstmed.ac.uk (M.G.M. Gomes), theolis@bahia.fiocruz.br (T.C.B. Bessa), randrade@ufba.br (R.F.S. Andrade), mauricio.barreto@fiocruz.br (M.L. Barreto).

https://doi.org/10.1016/j.tube.2020.102010
Received 22 May 2020; Received in revised form 30 September 2020; Accepted 12 October 2020
Available online 24 October 2020

(*continued*)

1. Background

Tuberculosis (TB) has long been a major public health concern worldwide. In 2017 alone, an estimated 10 million people developed TB, and 1.6 million died from the disease, including 300,000 deaths among human immunodeficiency virus (HIV)-positive individuals [1]. Despite an annual reduction of around 2% in the overall incidence of TB, this disease remains a significant public health issue, is one of the top ten causes of death and the leading cause from a single infectious agent, above HIV/AIDS [1].

TB control is largely based on the early diagnosis and treatment of TB cases in conjunction with the investigation of household contacts, for whom chemoprophylaxis can be provided. The early diagnosis and treatment of TB cases contribute for the interruption of the transmission chain, mainly within the household environment and particularly in individuals under 15 years of age [2]. On the other hand, in adult populations in large urban areas, contact beyond the household setting (i.e. extradomiciliary contact) plays an important role in the maintenance and transmission of TB [3–8]. Epidemiologic studies often give different or diverging results and have not consistently shown the role of extra domiciliary transmission. Many studies have shown that household contacts of individual with TB had higher risk of *M. tuberculosis* infection and disease than individuals in the general population [9–11]. Conversely, in Africa and South America, investigators in several studies have found that more than 60% of total transmission was attributable to public transport or community exposures [8,12–15].

Complex network analysis (NA) has been used in infectious disease epidemiology to evaluate connections between individuals (or groups of individuals) in order to understand the propagation and dynamics of disease [16]. In NA, individuals are represented by network nodes and disease transmission can only occur between two individuals who are linked by a network connection or edge. NA relies on the assumption that individuals differ in their social interactions, which is dependent not only on risk factors, such as socioeconomic level, education and migration, but also on how individuals are physically and socially integrated within communities. The extent of the influence of these differences and characteristics can be assessed using NA parameters in the same way that epidemiological analysis relies on risk measures [17,18].

NA allows for the construction of dynamic models to study the transmission of diseases, such as TB, particularly in the case of adult populations in urban centers whose extradomiciliary contacts are relevant. In such settings, the contact networks of healthy individuals, patients and carriers overlap, creating an extremely complex pattern of transmission [12]. Using NA, the most relevant nodes responsible for transmission can be identified and, based on connections within a given network, it is possible to predict which nodes are likely to be infected and to characterize spatial relationships. As subgroups of TB patients and contacts converge, specific collections of nodes can be selected for screening prioritization and areas of geographical confluence outside the household can be addressed.

This work aims to present an overview of the methodological and operational aspects of social network analysis, using TB as an example. We describe the use of *M. tuberculosis* genotyping results in conjunction with a comparative analysis between the social network patterns of TB cases and controls in different places of extradomiciliary social interaction, such as at school or work environments, to provide information on extradomiciliary TB transmission in a moderately endemic Brazilian city. We discuss how the information obtained from molecular epidemiology approaches can be integrated in NA to improve the overall robustness of the results.

2. Methods

2.1. Data structure

2.1.1. Study area, population and design

The study was conducted in the city of Salvador (the capital of the state of Bahia, located in northeastern Brazil), with a population of 2.7 million inhabitants in 2012 [10]. This city covers a total area of 706.8 km^2. The incidence of TB in Salvador was 70.64/100,000 in 2010 [11]. TB diagnosis and treatment are covered at no cost by the Brazilian Unified Health System.

The present case-control study was conducted between August 2008 and April 2010 and considered incident TB cases paired to controls by age (with a variation of ±5 years) and sex at a ratio of 1:1. The study population was composed of 717 cases and 717 controls, providing 95% power, 5% significance level, Odds Ratio of 2.0, frequency of exposure in controls 10%, considering 20% losses. The sample size calculation resulted in at least 278 cases with their respective controls. The 717 cases and 717 controls were recruited from five healthcare institutions with the highest frequency of tuberculosis.

2.2. Definition of cases and controls

Participants were selected from among individuals who presented with respiratory symptoms at the main municipal health care institutions responsible for TB referral: the outpatient clinics of three reference hospitals and six primary health care clinics in the city of Salvador, Bahia in northeastern Brazil. All candidates considered for this study were interviewed by the research team and underwent a sputum test.

Individuals with no previous history of TB, but with a positive sputum test subsequently confirmed by culture for *M. tuberculosis*, were considered as cases, while those with negative sputum tests, negative radiological test, no other signs or symptoms suggesting a pulmonary TB diagnosis, and no positive TB test for the next five years were considered as controls in accordance with the pairing criteria. These patients with respiratory symptoms were chosen as controls only after negative sputum tests, a TB nurse or clinician had discharged them as a no TB patient, and no positive TB test for the next five years. Controls were paired to cases with respect to sex, age, and reported current residence in Salvador. All included individuals signed an informed consent.

The inclusion criteria for this study consisted in individuals over 15 years of age, currently residing in Salvador and no history of HIV.

2.3. Data collection

Data were collected between August 2008 and April 2010 by a team of trained auxiliary nursing working under the supervision of senior nurses. Interviewers used a standardized questionnaire to collect clinical, sociodemographic and epidemiological data (supplementary material). In addition, questions were asked regarding the individual's life history during the preceding five years, focused on the places in which the individuals lived, studied, worked and participated in leisure activities. The participants' addresses were transcribed according to the answers obtained in the questionnaires. There was no corroboration of

these locations by the physical visits by the study investigators. The addresses were georeferenced for the formation of the home, work, study and leisure network. A manual containing detailed instructions on how to complete the questionnaires was provided to the interviewers. Monthly meetings were held with the supervisors and interviewers to review the completed questionnaires and to resolve any issues that arose during the process.

After conducting interviews, sputum samples were obtained by expectoration. The material was collected in sterile vials, properly stored under refrigerated conditions and sent to the state reference laboratory for microbiological confirmation of TB.

2.4. Study variables

The dependent variable was TB and independent variables consisted of demographic variables (sex, age, marital status, ethnicity), socio-economic variables (monthly family income, education level), history of TB contact (any previously known contact with an individual with TB), chronic disease (diabetes), domiciliary overcrowding and genotyping. Variables regarding lifestyle habits were drug use, smoking and alcohol consumption.

2.5. Genotyping

All sputum samples were processed using Petroff's method and cultured in Löwenstein-Jensen media (Becton-Dickinson, Palo Alto, CA, USA) at 37 °C for up to eight weeks. All mycobacterial isolates were submitted to phenotypic characterization methods and biochemical testing to identify *M. tuberculosis*. One or two loops of bacterial mass were frozen in 2 mL of Sauton medium (Becton-Dickinson, Palo Alto, CA) supplemented with 20% glycerol and kept at −70 °C. For DNA extraction, aliquots were defrosted and reactivated in Löwenstein-Jensen media at 37 °C for up to eight weeks. DNA was then extracted in accordance with the method used by van Embden et al. [12].

The IS6110 restriction fragment length polymorphism (RFLP) analysis was successfully performed in part of the isolates according to the method of van Embden et al. [12].

Microbead-based *spoligotyping* was performed according to the technique established by Cowan et al. and modified by Kamerbeek et al. to assess DNA polymorphisms within the direct repeat (DR) locus of *M. tuberculosis* [19,20]. Samples were consecutively assessed in a particulate solid phase fluorometer (Luminex, Austin, TX, USA). Cultures of the reference strains *M. tuberculosis* H37Rv and *M. bovis* were used as positive controls, while distilled water was used as a negative control. For each isolate, the median number of relative fluorescence units (MRFU) obtained for each spacer in each isolate was divided by the MRFU obtained for the negative control sample. A spacer was considered to be present in the genome of a given isolate when this ratio exceeded 5.0. Each 43-digit spacers profile was compared to the available records contained in the SITVIT database (http://www.pasteur-guadaloupe.fr:0881/SITVITDemo/).

Additionally, a set of 90 single nucleotide polymorphisms (SNP) was selected from a previously described pool of polymorphisms that were identified through genomic comparisons of the *Mycobacterium tuberculosis* H37Rv, CDC1551, and *M. bovis* AF2122/97 genomes [16–19] as detailed in Lopes et al. [20]. For this, primers were designed using the gene information and characterization retrieved from the GenBank Overview (http://www.ncbi.nlm.nih.gov/genbank/) and Pubmed (www.ncbi.nlm.nih.gov) databases. Most samples were genotyped with respect to the SNPs by following automated protocols using primer extension chemistry on a Sequenom MassArray platform [21]. Genomic sequences were amplified by multiplex PCR and the reaction product was treated with shrimp alkaline phosphatase and used for allele-specific primer extension in accordance with the MassEXTEND protocol. The reaction mixture was spotted onto a SpectroCHIP microarray and subjected to MALDI-TOF mass spectrometry. The genotype calls were assigned using SpectroTYPER software in accordance with SNP-specific peaks. For quality control, the genotyping process used *M. tuberculosis* strains EAS054, H37Rv, Haarlem, F11, C and CDC1551, all of which have curated and publicly available genomes.

2.6. Network analysis

In the context of social networks, the nodes and the edges between the nodes represent, respectively, the social actors (individuals) and the connections between them. Such connections depend on the type of system that is being evaluated: friendships, co-authors of a book or an article, co-actors in a movie or a play, etc. Other social networks are based on the relationships between people and objects or places they had in common, giving rise to bipartite networks. This was the case of the systems investigated in this study, in which people and places constitute a set of nodes, yet people are initially only connected to the places they attend. Based on these bipartite networks, a non-weighted person-person network (PP-network) can be constructed, consisting of only a single type of node (person). The nodes become connected if and only if the two persons frequent the same place, as shown in Fig. 1. Accordingly, PP-networks present a reasonable option for studying the role of extradomiciliary contacts in cases of infectious diseases, based on the links formed by shared places over a given time interval, as previously reported in the literature [9]. Supplementary material provides a description of the software program, referred to as GraphTube, which was developed to generate the networks.

In supplementary material, we also list the network basic concepts as well as the basic connectivity ratio. There, we also call the attention to the way we measure the dissimilarity between networks, which is based on the topological distance that was introduced by some of us and other collaborators in Refs. [22]. It depends on the neighborhood matrix, another representation of a graph, usually represented by adjacency matrix M (Figure SM-I).

2.6.1. Identifying/searching for the optimal PP network

The initial possibility of constructing PP networks is based on the coincidence of the places both people attend. However, it is possible to consider a criterion to link two people. The basic assumption for the inclusion of a link in a PP network is that the closer two people live or work to each other, the greater the probability that *M. tuberculosis* infection could be shared via epidemiological linkage.

The technique used here takes advantage of data pertaining to geo-referenced places of residence, school and workplace. Using this information, we evaluate the geodesic matrix G, in which the elements correspond to the shortest path connecting a geo-referenced attribute (home, school, workplace) of subjects i and j. In this work, the geodesic matrix G is the basic source of information for the geodesic distance, which is a proxy for the geographic distance. From now on, these terms indicate the same measure. G allows us to obtain a one-parameter family of geodesic networks (GN), which depends on the threshold geographical distance r. Each element in this family is defined by a network adjacency matrix $g(r)$, for which the elements $g_{i,j}(r)$ are defined as

$$g_{i,j}(r) = \begin{cases} 1, & if\ G_{i,j} \le r \\ 0, & if\ G_{i,j} > r \end{cases} \quad (1)$$

The selected matrix elements establish a direct link only between subjects i and j and only if the geodesic distance between them is smaller than r. For a given GN, we estimate a value $r = r^*$ so that the network g (r^*) would be more likely to provide useful information on the actual contact network [23]. To evaluate r^*, we use the network dissimilarity measure Δ defined in supplementary material.

The peaks of $\Delta(r, r + \delta r)$ indicate the values of r in which important changes in the network structure are introduced. We propose to consider that such peaks correspond to critical r^* values, once they identify when key links are introduced into the network, allowing for disease transmission in the whole area of study. Indeed, we select these critical r^*

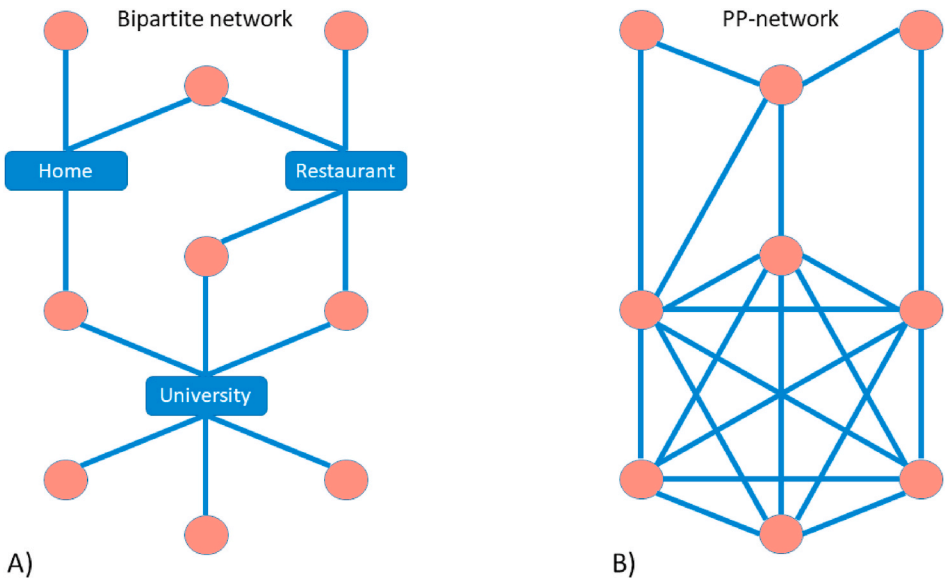


Fig. 1. An example of a social network analyzed in the present study: (A) A bipartite network in which the circular nodes represent people and the rectangular nodes represent places. (B) Person-person (PP-) network generated from the bipartite network shown in (A), assuming the existence of a link when both people attend the same place.

values as optimal values due to the reasons as follows: if r is small, it is very likely that $G_{ij} > r$ for almost all (i,j) pairs. In this situation, $M(r)$ leads to a network of essentially dissembled nodes, with only a few pairs of coupled nodes. On the other hand, if r is chosen to be of the order of the size of the town, almost all pairs of nodes (i,j) will satisfy $G_{ij} < r$, and the corresponding network will be very close to a completely connected graph. In the former case, the network fails to include links that are important in the understanding of disease transmission, while the latter extreme establishes connections among individuals with a very low likelihood of sharing a common place. The evaluation of r^*, whose value lies between these two extremes, is not merely a theoretical estimation of a critical geographical distance between individuals for social contacts, but is rather derived directly from the collected data.

However if the largest cluster presents an onion-type growth when the value of r increases, the best strategy is to observe both the peak of one of the measures; besides $\Delta(r,r+\delta r)$, the network shortest path $<\ell>(r) = \ell(r)$ or the diameter $D(r)$, that also exhibit peaks and the size of the largest cluster. If it is not large enough for r^* corresponding to the peak, the best choice to guarantee the network analysis (large enough number of nodes) is to assume $r > r^*$, but r should be close to r^*.

Once we have established an approach to set up an optimal geographical distance threshold for TB transmission, this can be used to construct case (CA) and control (CO) networks for use in other methods of characterization. In other words, these optimal networks will serve as input data for subsequent methods described below.

2.7. Network dissimilarity at critical threshold

Since it is important to compare the topology of CA (cases) and CO (control) PP-networks, the minimum values of dissimilarity between critical networks, Δ_{min}(CA,CO), were calculated according to the criteria established in sub-section 2.3.1. The CA and CO PP networks can also be compared by the used data: residence (R), school (S) or work (W) assuming the same number of nodes (see more details in Supplementary Material for different number of nodes).

To establish how different CA and CO networks are from each other, we evaluate the dissimilarity between one of the chosen networks (for instance, the CA network) and a random network generated from a large number of paired control samples.

2.8. Distributions of indices for critical networks

This method characterizes different features of the topological structure of each network. Essentially, the idea is to apply different databases to construct networks and to obtain basic network indices, whose concepts are listed in supplementary material, used to test hypotheses related to the structure of the two person-person networks: the case network and the control network. This analysis can reveal differences in relevant factors, such as mobility, between cases and controls. The network indices used were: node degree $< k > = k$, clustering coefficient $< c > = c$, eccentricity $< \varepsilon > = \varepsilon$, closeness centrality C_C and betweenness centrality C_B. These indices can be considered as the elements of a set that characterizes each of the groups assessed (cases/controls). A Wilcoxon test, paired by age and sex, was used to compare these indices.

2.8.1. Identifying patterns in the complete network of cases and controls

Considering now the larger set of all the individuals present in both CA and CO databases, two methodologies are proposed to characterize the neighborhoods of individuals in the CA and CO sets: the mean geographical distance between individuals and the connectivity neighborhood ratio.

Student's t-test was then used to evaluate the difference in the average between groups.

2.9. The mean geographical distance

To obtain geographical distance information from a given network, the geodesic matrix G was used. In the present study, this matrix can be grouped in order to display averages in the following groups: the average geographical distance between cases (CA-CA), between cases and controls (CA-CO) and between controls (CO–CO). The geodesics are contained in each pair of nodes.

2.10. Connectivity neighborhood ratio

In general, in a two-type node network, the probability of one node type being in the neighborhood of another type must be determined. For instance, in the analysis of a PP-network, it is important to know whether a CA individual is a neighbor of a CO or another CA so that

estimation can be obtained of the disease transmission within a social network.

The neighborhood ratio $R_{s \to t}$ estimates the fraction of nodes of type *s* (source) that are neighbors of nodes of type *t* (target) in a PP-network (see its expression for calculation in Supplementary Material). The neighborhood ratio $R_{s \to t}$ is strongly dependent on the ratio between the number of type *s* and type *t* nodes in a given network. To account for the effect of sampling sets of different sizes, the probability of significance $Z_{s \to t}$ measures how greater is the ratio $R_{s \to t}$ in relation to the same fraction expected for a random network with the same number of nodes and edges (see its expression also in Supplementary Material), maintaining an identical proportion of each node type as in the original network.

2.11. Integrating genotyping and social data

Important information on the TB transmission is provided by both genotyping and social data, which are collected and analyzed by completely different methods. Thus, it is important to investigate whether, by linking the information on these two data sets of different origins, it is possible to get a more consistent picture of the transmission process. In this work we explored two direct ways to combine these two sources of information, although it is well possible that other strategies for the same purpose can be developed.

We identified groups of strains according to their degree of similarity, using a dendrogram (data not shown). The sensitivity of the genotyping method is an important issue here, as increasing the sensitivity may make it more difficult to recognize isolates that would be similar enough to be considered to belong to the same group, while decreasing the sensitivity may blur the definition of groups that would have relevant similarity. We have combined the SNP patterns to the *spoligotype* patterns to perform the evaluations presented here, in an attempt to balance the measures of similarity between the isolates and the recognition of relevant groups.

Then, we compared the average geographic distance *r* between the residences of pairs of subjects represented in each group with the corresponding average taken over all pairs of subjects in the study. Next, we evaluated the changes in the average distance *r* within each group, caused by evaluation of the minimal distance between subjects with the inclusion of the information on the study and work places r_{RWS}. The two proposed comparisons might allow detecting a possible correlation involving the degree of similarity from genotypic information and the average distance between the places frequented by the subjects.

3. Results

3.1. Network analysis

Here we present actual results produced by our NA for the data set described in Section II.1. Only one case or control per household was identified. In Fig. 2, we present some network measures of CA and CO networks as a function of geographical distance r using the residence data. Fig. 2a, b, 2c and 2d show the dependence of $N_c(r)$, $\Delta(r,r+\delta r)$, $\ell(r)$ and $D(r)$ as a function of *r*. Fig. 2a reveals that the largest cluster $N_c(r)$ presents an onion-type growth with the geographical distance r. Fig. 2b correspond to network dissimilarity $\Delta(r,r+\delta r)$, revealing that the critical value is $r^* = 200$ m for both CA and CO networks. Besides $\Delta(r,r+\delta r)$, other measures, such as the network shortest path $<\ell>(r) = \ell(r)$ and the diameter $D(r)$, also exhibit peaks. Due to the onion-type growth of N_c with r, the optimal distance for our analysis is r = 300 mts, slightly

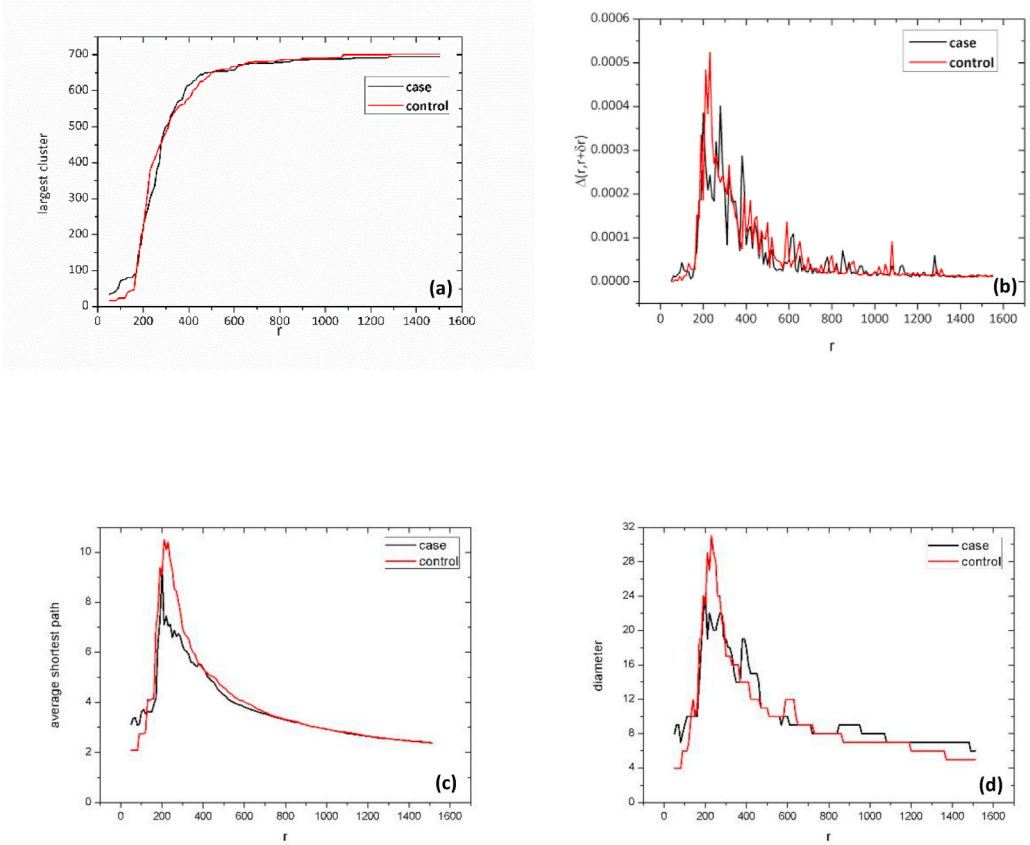


Fig. 2. Critical network analysis of tuberculosis based on residence network in accordance with the distance *r*: (a) size of the largest cluster; (b) network dissimilarity between *r* and $r + \delta r$; (c) average shortest path; (d) diameter. The critical threshold is $r^* = 200$ m while, for the analyzed network, the used value $r = 300$ m is slightly larger than r^* For $r = 300$ m there are 400 nodes in the largest cluster.

higher than r* = 200 mts, for which the networks have approximately 400 nodes. Fig. 3a–c shows respectively, $N_c(r)$ and $\ell(r)$ as a function of r, of CA and CO networks based on extradomiciliary data (School or Work taken together). The obtained critical threshold value is r* = 300 for extradomiciliary data (Fig. 3c); for the same reason, the analyzed network uses the value r = 400 mts, slightly higher than r* = 300 mts. Finally, setting up the networks considering both household (residence) and extradomiciliary (school or workplace) places together, the values of r* are 100 mts and 200 mts, respectively, for CA and CO networks as it is shown in Fig. 3d; although the onion type growth of N_c with r, the largest clusters of CA and CO critical networks have a significant number of nodes (Fig. 3b).

In the following step, we evaluated the network dissimilarity between the random versions of the CO and CA networks and then compared this to the network dissimilarity between the actual CA and CO networks (Fig. 4). In this case, the actual networks were more similar to each other than their random versions; moreover, the complete (residence + school + workplace) CA and CO networks were less similar to each other than residence CA and CO networks and extradomiciliary CA and CO networks.

Since the CA and CO networks restricted to the residence data are constructed more securely due to the fact that the addresses of all participants are registered, we selected those networks to show the network indices, whose distributions were compared using the nonparametric Mann-Whitney test [24], as these did not meet the Kolmogorov-Smirnov criteria for normality. The critical indices of the networks for TB cases and controls, based on residence data, are shown in Table 1. Our results show that networks of cases and controls have significantly ($p < 0.05$) different index distributions for eccentricity and closeness centrality, with higher ε values and lower C_C values for the control distributions.

We also calculate the mean geographic minimal distance between

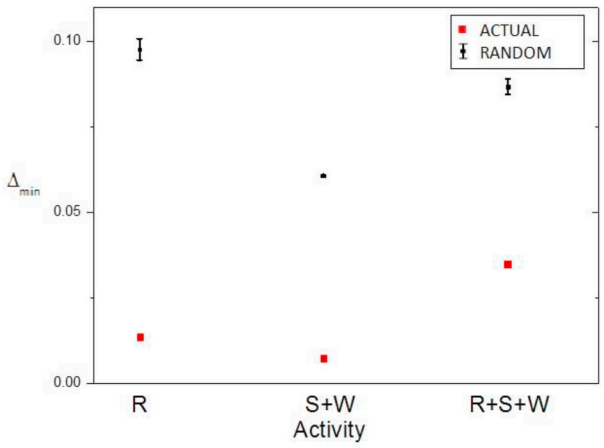


Fig. 4. The network dissimilarity between case (CA) and control (CO) critical networks for different data (Δmin): residence data ($r^* = 300$ m), extradomiciliary data (workplace + school) ($r^* = 300$ m), all databases ($r^* = 200$ m). Red points correspond to the actual CA and CO networks. The minimal value of the network dissimilarities of the random versions of CA and CO networks are represented by black points with error bars. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

groups as described in sub-section 2.3.4 and defined in supplementary material. These results, shown in Table 2, confirm the previously obtained results of the topological indices in the residence network.

Finally, the results for the connectivity neighborhood ratio were based on the neighborhood ratio R and the probability of significance Z

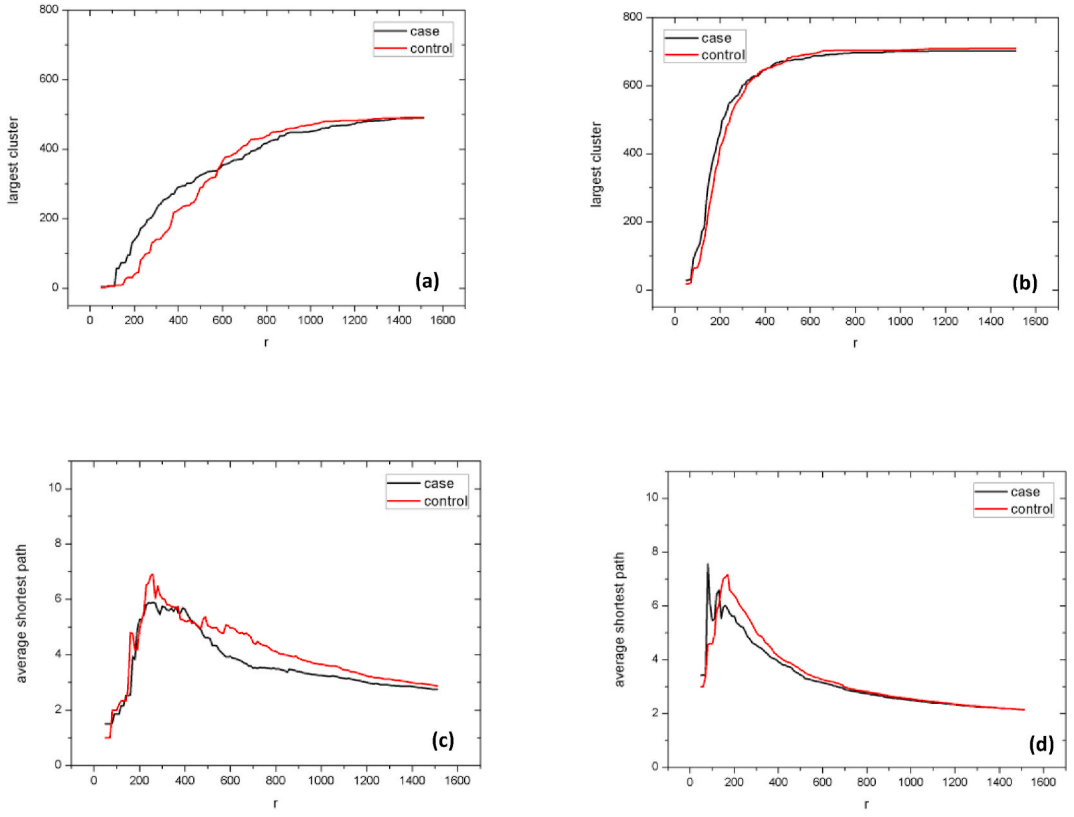


Fig. 3. Results from the tuberculosis critical network analysis for CA and CO, for extradomiciliary data and all databases: (a) $N(r)$ x r with extradomiciliary data (workplace + school), for which the critical threshold is $r^* = 300$ m, and the used value of the analyzed network is $r = 400$ m a value slightly larger than r^*; (b) $N(r)$ x r for all databases, for which, $r^* = 100$ m and r* = 200 m for CA and CO critical networks respectively; (c) $\ell(r)$ x r with same extradomiciliary data as in (a); (d) $\ell(r)$ x r for all databases as in (b).

Table 1
Indices corresponding to the critical residence network of tuberculosis (r = 300 mts).

Indices	Case		Control		
	Mean	Median (p_{25};p_{75})	Mean	Median (p_{25};p_{75})	*p-value
K	6.04	4.0(2.0–8.0)	5.86	4.0(2.0–8.0)	0.799
E	7.97	9.0(8.0–10.0)	8.46	10.0(9.0–11.0)	0.001
C_C	0.25	0.21(0.16–0.25)	0.23	0.20(0.16–0.23)	0.017
C_B	0.003	0.002(0.0–0.03)	0.003	0.0(0.0–0.004)	0.447
C	0.59	0.66(0.33–0.96)	0.57	0.61(0.33–0.931)	0.312

k: node degree; ε: eccentricity; C_C: closeness centrality; C_B: between ness centrality; c: clustering coefficient.

Table 2
Mean minimal geographical distance between groups in residence locations. All values are significantly different using a *t*-student test ($p < 0.05$).

Group	Distance (m)
CA-CA	6568
CA-CO	6684
CO–CO	6713

CA: case; CO: control.

for the neighborhood ratio R for the residence data. They were calculated taking into consideration all the neighborhood possibilities: case-case (CA-CA), case-control (CA-CO), control-case (CO-CA) and control-control (CO–CO) (Table 3). It is important to observe that just one value average distance between case and control subjects is reported in Table 2, as this is a symmetric function. However, the same does not hold for the values of R for CA-CO and CO-CA in Table 3. Indeed, in this case the role played by two types of nodes in Eq. (3) of the Supplementary Material is not symmetric. Finally, we recall that, as the neighborhood ratio R was obtained through the comparison to random networks, the reported values for Z are equivalent to the corresponding p-values.

3.2. Genotyping

We have successfully determined the spoligotype and SNP genetic profiles of 342 isolates. The genetic profiles obtained allowed for the identification of the Spoligotype International Type of 299 isolates, while 43 isolates consisted of orphan patterns. RFLP profiles were additionally obtained for 319 isolates. We obtained ten groups of strains with at least 85% of similarity, twice as many groups with at least 88% of similarity and 86 groups at the 98% similarity threshold. The diversity of RFLP patterns was evident even among strains pertaining to 100% genetically similar groups when combining *spoligotyping* and SNP data, e.g. among isolates belonging to LAM 9 SIT 42. This is expected, as the patients analyzed have no direct epidemiological connections. More refined genotyping techniques can be used to recognize similarity among strains in circulation in replacement of the techniques here employed, e.g. core genome Multilocus Sequence Typing (cg-MLST) [25].

Table 3
R and Z values for different combinations of groups in the residence database.

Group	R	Z^a
CA-CA	0.536	0.000
CA-CO	0.464	1.000
CO-CA	0.477	1.000
CO–CO	0.523	0.003

CA: case; CO: control; R: neighborhood ratio; Z: the probability of significance.
[a] For the Z calculation, 10^3 independent randomizations were performed.

3.3. Integration of genotyping and social data

The average geographic distance r between the residences of subjects within the same genetic groups did not differ from the all-pairs average (Table 4), which can be demonstrated by the overlapping confidence intervals. Therefore, patients with genetically similar isolates were not found to be clustered geographically, when considering only the distances between their residences.

The result for the second procedure reveals a strong reduction from the average distance r to r_{RWS} (that integrates the information of study and work places) in the evaluation of intra-groups average distances (Fig. 5). This result would be consistent with tracking possible transmission pathways involving encounters in closer geographic spaces other than the household, as the proximity between individuals from whom the most similar isolates were retrieved would be thereby recognized. Yet, from our analyses we demonstrate that the geographic distance r_{RWS} accounts for a very limited proportion of the observed variability in the genotypic similarity as measured in this work, as would be desirable to enhance the applicability of our framework. All measures of distance - r, r_{RWS} (Fig. 6) and the difference r-r_{RWS} (data not shown) do not correlate with genotypic similarity.

4. Discussion

The network analysis performed here provides a comparative scenario between the networks of a population of TB cases and a population of healthy controls, determined by the optimal network dissimilarity that set up the critical network. To start with, the first important point is to determine the optimal network for the subsequent comparative analysis of CA and CO networks. The critical network analysis for CA and CO using different data show several peaks on some network measures that reveal the existence of relevant differences in CA and CO network structures.

The presence of just one peak or a larger number of peaks on some network measures depends on the actual data collected and on the geographic distribution of individuals within the community under consideration. In previous studies [23,26], the most efficient network for detecting the modular behavior of a network was found by selecting the r^* that corresponds to the maximal peak.

In our study, for both CA and CO networks, $r^* = 200$ mts using residence data meanwhile $r^* = 300$ mts for the extradomiciliary database. Analogously the optimal distances for both residence and extradomiciliary data are chosen $r = 300$ mts, and $r = 400$ mts respectively since the size of the largest clusters were not large enough for r^* corresponding to their peaks. That analysis reveals that there is no difference between the optimal distances for CA and CO networks, but the optimal distances are smaller for residence data than for extradomiciliary data. On the other side, taking into account all databases (Residence, School or Workplace), the smaller value of critical distance for CA network than for CO network pointing out that TB cases have a pattern behavior slightly different from controls. Note that the critical distances coincide to the optimal distances since the size of the largest cluster grows faster with the distance r; it is reasonable because considering all places, there are more links between the individuals. Finally, it is important to mention that the description of the complete CA network and CO network, taking into account all data provide a general signature of the dynamics of TB cases and controls, respectively.

The difference between CA and CO complete networks have also revealed in the analysis of the network dissimilarity (Fig. 4) since it is higher than its difference for residence data and extradomiciliary data. Moreover, the difference between the random version of CA and CO networks and the actual networks for each data (residence,

Table 4
Mean distance *r*(m) between residences of pairs of patients with strains belonging to the major genetic subgroups identified.

Genetic subgroup	N of patients	N of P–P pairs	*r* (m)	Standard deviation	95% CI
I A 3 j	24	276	7823.05	4439.48	7297–8349
I B 4 r	10	45	7602.45	4509.30	6248–8957
III F 12 al	16	120	11940.27	8797.54	10,350–13530
III F 12 na	48	1128	7827.39	4876.28	7543–8112
III F 12 ap	14	91	6537.86	3834.36	5739–7336
III F 12 at	24	276	6171.54	3258.49	5785–6558
III F 12 aw	10	45	7610.06	4568.85	6237–8983
III F 13 br	25	300	8198.94	5497.98	7574–8824
All genotyped	305	46,360	7597.84	5000.58	7552–7643

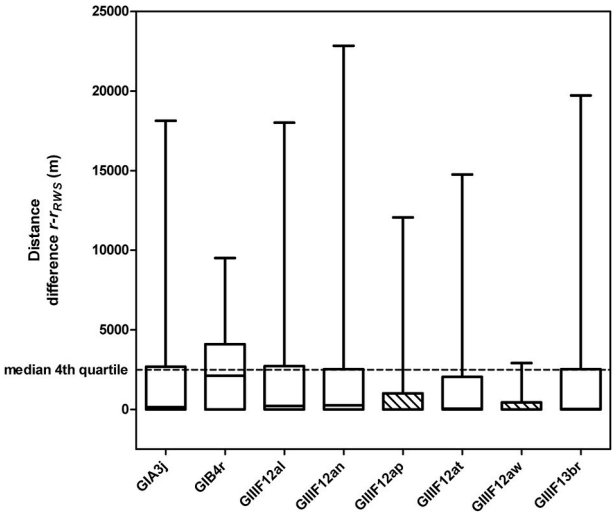


Fig. 5. Distance difference r-r_{RWS} considering the pairs of patients with strains belonging to the major genetic subgroups identified. The dashed horizontal line indicates the median value for the lower limit of the fourth topper quartile, balanced by the number of individuals per group. Subgroups GIIIF12ap and GIIIF12aw have shown little modification of the distance values when considering r_{RWS} and were not considered for further analyses.

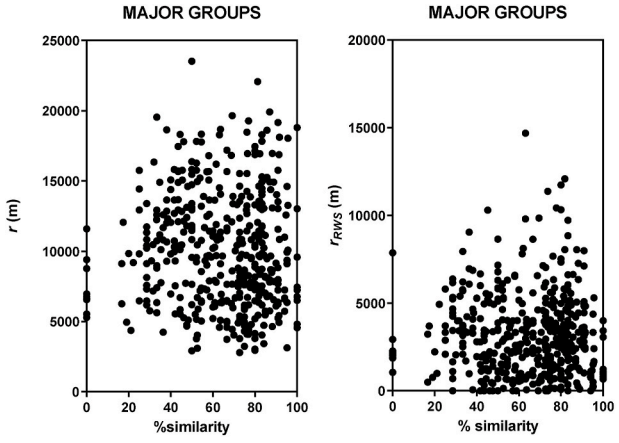
extradomiciliary and all database) reveals the non-random character of the actual networks.

In order to investigate comparatively how topologically close the CA nodes are from CO nodes, we calculate the connectivity neighborhood ratio (SM-2) whose values indicate a homophily behavior: CA are significantly more likely to be connected with CA and CO with CO (Table 3). A possible explanation for this behavior is that places of residence shared by cases are geographically closer than places of residence shared by controls. The results of the mean geographic minimal distance between groups confirm that hypothesis, indicated by the indices, that cases live in closer proximity than controls.

Another relevant question is to get some information about connectivity neighborhood ratio, based on the network with TB cases and controls together. The results for the neighborhood ratio R and its probability of significance Z for the residence data show that cases have significantly more places in common with other cases than with controls, while controls correspondingly share more places with each other. This means that the sharing network involves clusters of cases and clusters of controls.

The results concerning the genotyping allowed the identification of eight relevant groups with high similarity distributed in Salvador, that involved individuals widely dispersed in the city, when considering the places of residence, for whom possible epidemiologic links were not evident.

The association between genotyping and social data used in the network analysis produced some interesting results. However, it was not possible to assign a clear-cut correlation between the reduction in the average distance between the places visited by the subjects with highly similar isolates with the corresponding all-pair average. The clear reduction in the average distance when the average is taken over the minimal distance based on all data for residence, study and work places was also insufficient to establish a correlation with the genotypic similarity. This may indicate that other relevant information regarding places visited by the studied subjects may not have been captured. For instance, Sacchi et al. [27] show a relevant contribution of the prison population to the transmission of tuberculosis in a low endemicity setting. From the data collection design, we followed it was not possible



	r	rRWS
Spearman r	-0.07146	-0.03099
95% confidence interval	-0.1646 to 0.02293	-0.1249 to 0.06343
P value (two-tailed)	0.1263	0.5077

Fig. 6. Correlation between the distance measures r, r_{RWS} and the degree of similarity among the strains isolated from the individuals in each pair, considering only individuals that had strains from the major genetic subgroups identified and for which the distance difference r-r_{RWS} was superior to the mean value for the lower limit of the fourth topper quartile shown in Fig. 6 (corresponding to individuals with genotypically similar strains that were found to be in closer proximity when considering the geographic information of work and study in addition to the residence).

to obtain information regarding previous incarceration or contact with previously incarcerated subjects, or identify any other overcrowded places that were putatively shared in an intensive way by the study subjects.

5. Concluding remarks

Network analysis can aid in the construction of dynamic models to study TB transmission, particularly in the case of adult populations in urban centers whose extra-household contacts are relevant. NA parameters help to describe, model and investigate the difference between the TB cases network (CA network) and control network (CO network) using the optimal networks based on the critical distance. The use of networks in the study of TB dynamics and TB natural history can also further the existing knowledge about the role of extradomiciliary transmission of TB, by compiling information regarding contacts (either healthy or infected) together with information from TB index cases. The genotypic data make it possible to identify and characterize several clusters with high similarity, which provides a measure of the relative importance of the modelled connections for disease occurrence and transmission. We believe that this innovative approach can produce knowledge useful for the guidance of epidemiological surveillance aimed at improving tuberculosis control.

Acknowledgements

The authors thank Dr Martha Oliveira for conducting *spoligotyping*, and Dr Carlos Penha for his contribution in SNP. STRP and RFSA were supported by the National Institute of Science and Technology – Complex Systems from CNPq - Brazil.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.tube.2020.102010.

References

[1] World Health Organization. Global tuberculosis report 2018. Geneva: World Health Organization; 2018. License: CC BY-NC-SA 3.0 IGO. Available from: https://apps.who.int/iris/bitstream/handle/10665/274453/9789241565646-eng.pdf?ua=1.

[2] Cohen T, Colijn C, Finklea B, Murray M. Exogenous re-infection and the dynamics of tuberculosis epidemics: local effects in a network model of transmission. J R Soc Interface 2007;4:523–31.

[3] Daley CL. Molecular epidemiology: a tool for understanding control of tuberculosis transmission. Clin Chest Med 2005;26:217–31 [vi].

[4] Classen CN, Warren R, Richardson M, Hauman JH, Gie RP, Ellis JH, et al. Impact of social interactions in the community on the transmission of tuberculosis in a high incidence area. Thorax 1999;54:136–40.

[5] Yaganehdoost A, Graviss EA, Ross MW, Adams GJ, Ramaswamy S, Wanger A, et al. Complex transmission dynamics of clonally related virulent Mycobacterium tuberculosis associated with barhopping by predominantly human immunodeficiency virus-positive gay men. J Infect Dis 1999;180:1245–51.

[6] Danon L, Ford AP, House T, Jewell CP, Keeling MJ, Roberts GO, et al. Networks and the epidemiology of infectious disease [internet]. Interdisciplinary perspectives on infectious diseases. Available from: https://www.hindawi.com/journals/ipid/2011/284909/; 2011.

[7] Hollm-Delgado M-G. Molecular epidemiology of tuberculosis transmission: contextualizing the evidence through social network theory. Soc Sci Med 2009;69:747–53.

[8] El-Sayed AM, Scarborough P, Seemann L, Galea S. Social network analysis and agent-based modeling in social epidemiology. Epidemiol Perspect Innovat 2012;9:1.

[9] Klovdahl AS, Graviss EA, Yaganehdoost A, Ross MW, Wanger A, Adams GJ, et al. Networks and tuberculosis: an undetected community outbreak involving public places. Soc Sci Med 2001;52:681–94.

[10] IBGE. Instituto Brasileiro de Geografia e Estatística [Internet]. Available from: http://www.ibge.gov.br/home/estatistica/economia/perfilmunic/2012/ [cited 2017 Jul 25].

[11] TabNet Win32 3.0: D.2.2 Taxa de incidência de tuberculose [Internet] [cited 2017 Jul 22]. Available from: http://tabnet.datasus.gov.br/cgi/tabcgi.exe?idb2012/d0202.def.

[12] van Embden JD, Cave MD, Crawford JT, Dale JW, Eisenach KD, Gicquel B, et al. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. J Clin Microbiol 1993;31:406–9.

[13] Cowan LS, Diem L, Brake MC, Crawford JT. Transfer of a *Mycobacterium tuberculosis* genotyping method, Spoligotyping, from a reverse line-blot hybridization, membrane-based assay to the Luminex multianalyte profiling system. J Clin Microbiol 2004;42:474–7.

[14] Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper S, et al. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. J Clin Microbiol 1997;35:907–14.

[15] Brudey K, Driscoll JR, Rigouts L, Prodinger WM, Gori A, Al-Hajoj SA, et al. *Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. BMC Microbiol 2006;6:23.

[16] Dos Vultos T, Mestre O, Rauzier J, Golec M, Rastogi N, Rasolofo V, et al. Evolution and diversity of clonal bacteria: the paradigm of *Mycobacterium tuberculosis*. PLoS ONE 2008;3:e1538.

[17] Filliol I, Motiwala AS, Cavatore M, Qi W, Hazbón MH, Bobadilla del Valle M, et al. Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set. J Bacteriol 2006;188:759–72.

[18] 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.

[19] Kasai H, Ezaki T, Harayama S. Differentiation of phylogenetically related slowly growing mycobacteria by their gyrB sequences. J Clin Microbiol 2000;38:301–8.

[20] Lopes JS, Marques I, Soares P, Nebenzahl-Guimaraes H, Costa J, Miranda A, et al. SNP typing reveals similarity in *Mycobacterium tuberculosis* genetic diversity between Portugal and Northeast Brazil. Infect Genet Evol 2013;18:238–46.

[21] Gabriel S, Ziaugra L, Tabbaa D. SNP genotyping using the Sequenom MassARRAY iPLEX platform. Curr Protoc Hum Genet 2009;60(2.12):1–18.

[22] Andrade RFS, Miranda JGV, Lobão TP. Neighborhood properties of complex networks. Phys Rev E - Stat Nonlinear Soft Matter Phys 2006;73:046101.

[23] Andrade RFS, Miranda JGV, Pinho STR, Lobão TP. Measuring distances between complex networks. Phys Lett 2008;372:5265–9.

[24] Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat 1947;18:50–60.

[25] Ghanem M, Wang L, Zhang Y, Edwards S, Lu A, Ley D, et al. Core genome Multilocus sequence typing (cgMLST): a standardized approach for molecular typing of *Mycoplasma gallisepticum*. J Clin Microbiol 2017;56:e01145–17.

[26] Andrade RFS, Rocha-Neto IC, Santos LBL, de Santana CN, Diniz MVC, Lobão TP, et al. Detecting network communities: an application to phylogenetic analysis. PLoS Comput Biol 2011;7:e1001131.

[27] Sacchi FPC, Praça RM, Tatara MB, Simonsen V, Ferrazoli L, Croda MG, Suffys PN, Ko AI, Andrews JR, Croda J. Prisons as reservoir for community transmission of tuberculosis, Brazil. Emerg Infect Dis 2015;21:452–5.