Infection, Genetics and Evolution 72 (2019) 31–43

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Macro-geographical specificities of the prevailing tuberculosis epidemic as seen through SITVIT2, an updated version of the *Mycobacterium tuberculosis* genotyping database

David Couvin*, Audrey David, Thierry Zozio, Nalin Rastogi*

WHO Supranational TB Reference Laboratory, Unité de la Tuberculose et des Mycobactéries, Institut Pasteur de Guadeloupe, Abymes, Guadeloupe, France

ARTICLE INFO

Keywords:
Mycobacterium tuberculosis
Genotyping
Database
Epidemiology
Phylogeography
Drug resistance

ABSTRACT

In order to provide a global overview of genotypic, epidemiologic, demographic, phylogeographical, and drug resistance characteristics related to the prevailing tuberculosis (TB) epidemic, we hereby report an update of the 6th version of the international genotyping database SITVIT2. We also make all the available information accessible through a dedicated website (available at http://www.pasteur-guadeloupe.fr:8081/SITVIT2). Thanks to the public release of SITVIT2 which is currently the largest international multimarker genotyping database with a compilation of 111,635 clinical isolates from 169 countries of patient origin (131 countries of isolation, representing 1032 cities), our major aim is to highlight macro- and micro-geographical cleavages and phylogeographical specificities of circulating *Mycobacterium tuberculosis* complex (MTBC) clones worldwide. For this purpose, we retained strains typed by the most commonly used PCR-based methodology for TB genotyping, i.e., spoligotyping based on the polymorphism of the direct repeat (DR) locus, 5-loci Exact Tandem Repeats (ETRs), and MIRU-VNTR minisatellites used in 12–, 15–, or 24–loci formats. We describe the SITVIT2 database and integrated online applications that permit to interrogate the database using easy drop-down menus to draw maps, graphics and tables versus a long list of parameters and variables available for individual clinical isolates (year and place of isolation, origin, sex, and age of patient, drug-resistance, etc.). Available tools further allow to generate phylogenetical snapshot of circulating strains as Lineage-specific WebLogos, as well as minimum spanning trees of their genotypes in conjunction with their geographical distribution, drug-resistance, demographic, and epidemiologic characteristics instantaneously; whereas online statistical analyses let a user to pinpoint phylogeographical specificities of circulating MTBC lineages and conclude on actual demographic trends. Available associated information on gender (n = 18,944), age (n = 16,968), drug resistance (n = 19,606), and HIV serology (n = 2673), allowed to draw some important conclusions on TB geo-epidemiology; e.g. a positive correlation exists between certain *Mycobacterium tuberculosis* lineages (such as CAS and Beijing) and drug resistance (p-value < .001), while other lineages (such as LAM, X, and BOV) are more frequently associated with HIV-positive serology (p-value < .001). Besides, availability of information on the year of isolation of strains (range 1759–2012), also allowed to make tentative correlations between drug resistance information and lineages – portraying probable evolution trends over time and space. To conclude, the present approach of geographical mapping of predominant clinical isolates of tubercle bacilli causing the bulk of the disease both at country and regional level in conjunction with epidemiologic and demographic characteristics allows to shed new light on TB geo-epidemiology in relation with the continued waves of peopling and human migration.

1. Introduction

The current century has seen a tremendous growth in the amount and complexity of microbiological data generated, and we are witnessing today the advent of a new era in collection and treatment of information accumulated. Like for many other infectious diseases, our knowledge on the understanding of TB as a disease as well as on the evolutionary aspects of its bacterium is improving day after day; nonetheless, the disease which is probably at least as old as humanity (Comas et al., 2013; Rothschild et al., 2001; Zink et al., 2003), remains a silent killer. *Mycobacterium tuberculosis* – the etiologic agent in question – is still spread worldwide killing millions of people each year, and represents a major threat for global public health.

* Corresponding authors.
E-mail addresses: dcouvin@pasteur-guadeloupe.fr (D. Couvin), nrastogi@pasteur-guadeloupe.fr (N. Rastogi).

https://doi.org/10.1016/j.meegid.2018.12.030
Received 1 October 2018; Received in revised form 23 December 2018; Accepted 25 December 2018
Available online 26 December 2018

Indeed, despite significant advances occurring in the field of molecular diagnosis and epidemiology of tuberculosis (TB), the recent global TB Report for 2017 from the World Health Organization (WHO) shows an overall picture of high burden of disease and slow progress to make any major headway in beating TB (WHO, 2017). Another deleterious development is the emergence of drug-resistant TB, particularly the multiple drug resistant (MDR) and extensively drug resistant (XDR)-TB, and more recently even the onset of totally drug resistant (TDR) or "pan-resistant" TB (Velayati et al., 2009), which are a significant threat to effective TB control. In such a context, the positive association of TB with HIV/AIDS constitutes a significant threat to public health in certain countries such as South Africa (Dheda et al., 2010).

Bearing in mind that TB epidemic seems to be related to poverty, promiscuity, poor living conditions, or other hard socio-demographic aspects (Farmer et al., 2006; Keshavjee et al., 2008), monitoring of drug resistant strains worldwide, especially in resource limited countries, is urgently needed for TB control. The WHO and its partners developed strategies such as the Green Light Committee to foster a better treatment of MDR and XDR-TB particularly in resource limited countries (Zignol et al., 2006; WHO 2017). Not abating the need of development of better diagnostic tests, effective treatment, and effective vaccine development, we are convinced that concerted and coordinated efforts and strategies are also needed in parallel to focus on evidence-based geographical mapping of predominant clones of tubercle bacilli causing the bulk of the disease both at country and regional level.

Regarding the methodology, we decided to work on creation, development and curation under an ongoing database program in development in our laboratory for last > 15 years (Sola et al., 2001; Filliol et al., 2002, 2003; Brudey et al., 2006). Right from the beginning, we selected robust PCR-based genotyping methods instead of fingerprinting using IS*6110*-Restriction Fragment Length Polymorphism (RFLP); even though it was considered as the gold standard in molecular epidemiology of TB in the nineties (van Embden et al., 1993). The complexity of this method and the labor-intensive methodology led to its replacement by easier yet more robust techniques by widely used PCR-based genotyping methods such as spoligotyping (Kamerbeek et al., 1997) and 12, 15, or 24-loci MIRU-VNTRs (Supply et al., 2006), which represent the current gold-standard for TB genotyping surveillance. Despite some limitations (Comas et al., 2009), their use to identify, track and characterize the *M. tuberculosis* clones has been validated in various studies (as reviewed in: Rastogi and Sola, 2007; García de Viedma et al., 2011; Jagielski et al., 2016). Other methods of choice for evolutionary and phylogenetical studies include Single Nucleotide Polymorphism (SNPs; Filliol et al., 2006), and Large Sequence Polymorphism (LSP) / Regions of Difference (RD; Gagneux et al., 2006).

Vis-à-vis the management of huge collections of data generated, microbiological databases are undeniably useful to make comparisons within isolates worldwide, and advances are obviously needed to offer improved analyses and innovations in the comprehension of biological data (Howe et al., 2008). In relation to the field of TB molecular or genomic epidemiology, several databases/web tools were developed over last 10 years to facilitate the comprehension and control of TB epidemics worldwide; examples include following online resources (list not exhaustive): SpolTools (http://spoltools.emi.unsw.edu.au/) (Reyes et al., 2008; Tang et al., 2008); MIRU-VNTRplus (http://www.miru-vntrplus.org/) (Allix-Béguec et al., 2008; Weniger et al., 2010); TB-Lineage (http://tbinsight.cs.rpi.edu/run_tb_lineage.html) (Shabbeer et al., 2012); SITVITWEB (http://www.pasteur-guadeloupe.fr:8081/SITVIT_ONLINE) (Demay et al., 2012) MycoDB.es (http://www.vigilanciasanitaria.es/mycodb/) (Rodriguez-Campos et al., 2012); Mbovis.org (http://www.mbovis.org/) (Smith and Upton, 2012); tbvar (http://genome.igib.res.in/tbvar/) (Joshi et al., 2014); InTB (http://www.evocell.org/inTB) (Soares et al., 2013); PolyTB (http://pathogenseq.lshtm.ac.uk) (Coll et al., 2014a); Genome-wide *Mycobacterium tuberculosis* variation (GMTV) database (http://mtb.dobzhanskycenter.org/) (Chernyaeva et al., 2014); CPLP-TB (http://cplp-tb.ff.ulisboa.pt/) focusing on TB in Portuguese-speaking countries (Perdigão et al., 2018); and SpolSimilaritySearch (http://www.pasteur-guadeloupe.fr:8081/SpolSimilaritySearch), which allows to compare and search similarities between spoligotypes (Couvin et al., 2017). Whenever feasible, these resources are optimally exploited in conjunction with the map library of WHO to conduct comparative analyses of TB in association with multiple parameters such as drug resistance, HIV serology, or incidence in respective countries (http://gamapserver.who.int/mapLibrary/app/searchResults.aspx).

In the present study, we describe and publicly release SITVIT2 which is a MySQL-based multimarker database containing genotyping information (spoligotyping, and 12, 15, or 24-loci MIRU-VNTRs) on 111,635 MTBC clinical isolates. In addition to the previously released version (Demay et al., 2012), it combines phylogenetic description of *M. tuberculosis* lineages in conjunction with detailed epidemiological analyses of circulating clones based on available demographic, epidemiologic, and drug resistance data of individual isolates. It therefore substantiates the need conveyed in a recent WHO handbook entitled "Understanding and using Tuberculosis Data" (WHO 2014), since in-depth geographical mapping of strains and identification of circulating clones within a broader sociodemographic and phylogeographic context will allow to pinpoint prevailing disparities, helping to better monitor, understand and control the TB epidemic worldwide.

2. Materials and Methods

2.1. Ethical statement and data collection

Collection and treatment of data was performed as described earlier for SITVITWEB (Demay et al., 2012). A total of 111,635 isolates were collected in the MySQL database SITVIT2 which is an updated version of earlier databases developed at Institut Pasteur de la Guadeloupe, namely: SpolDB1 to 4, and SITVITWEB. As for earlier versions, results on genotyping and related epidemiological and demographic data were collected from collaborating laboratories (details provided on the database website), and anonymized prior to entry in SITVIT2. Note that genotyping data contained in SITVIT2 database are from data received until May 2013; year of isolation of strains ranged from 1759 (from frozen cultures) – until 2012. Furthermore, most of the data contained in SITVIT2 was previously published as independent bilateral collaborations; and data entered after May 2013 will be made available in a future version of the database tentatively named SITVITEXTEND.

2.2. Genotyping

The genotyping data were limited to spoligotyping (Kamerbeek et al., 1997), and 12, 15 and 24-loci MIRU-VNTRs, (Supply et al., 2006), as well as a shorter 5-loci format of exact tandem repeats (ETRs). The available results correspond to 103,856 isolates with spoligotyping, and 23,919 with MIRU-VNTRs (split in: 12-loci MIRU $n = 19{,}311$; 15-loci MIRU $n = 4702$; 24-loci MIRU $n = 1672$; and 5-loci ETR $n = 10{,}294$). Throughout in this paper, the order of MIRU loci is as follows: 12-loci MIRU patterns: MIRU 2, 4, 10, 16, 20, 23, 24, 26, 27, 31, 39 and 40; 15-loci MIRU patterns: MIRU 4, 10, 16, 26, 31, and 40, ETR-A, ETR-C, QUB-11b, QUB-26, QUB-4156, Mtub04, Mtub21, Mtub30, and Mtub39; and 24-loci MIRU patterns: classical 12-loci pattern followed by ETR-A, ETR-B, ETR-C, QUB-11b, QUB-26, QUB-4156, Mtub04, Mtub21, Mtub29, Mtub30, Mtub34, and Mtub39. Spoligotype International Type (SIT), VNTR International Type (VIT), and 12–15 or 24-MIRU International Type (12–15 or 24-MIT) designate identical spoligotypes, 5-locus ETRs, 12–15 or 24-loci MIRU-VNTRs patterns respectively, that are shared by 2 or more patient isolates. Table 1 recapitulates the available genotyping information based on various molecular markers included in SITVIT2.

2.3. Statistical and bioinformatical analysis

STATA software version 12 was used for descriptive and univariate analyses. Comparisons between different genotyping, demographic and

Table 1
Quantitative description of genotyping markers contained in SITVIT2.

Genotyping marker	Number of distinct patterns	Number of strains	Number of International Types (IT)	Number of ITed strains[a]	Number of orphan strains	Percentage of orphan strains	Number of countries of isolation
Spoligotype	9658	103,856	3851	98,049	5807	5.6	131
5-locus ETRs (VNTR)	812	10,294	424	9539	755	7.3	39
12-loci MIRU	4773	19,311	1632	14,777	4534	23.5	53
15-loci MIRU	2327	4702	430	1827	2875	61.1	18
24-loci MIRU	876	1672	178	765	907	54.2	17

[a] This heading means number of strains having an IT number.

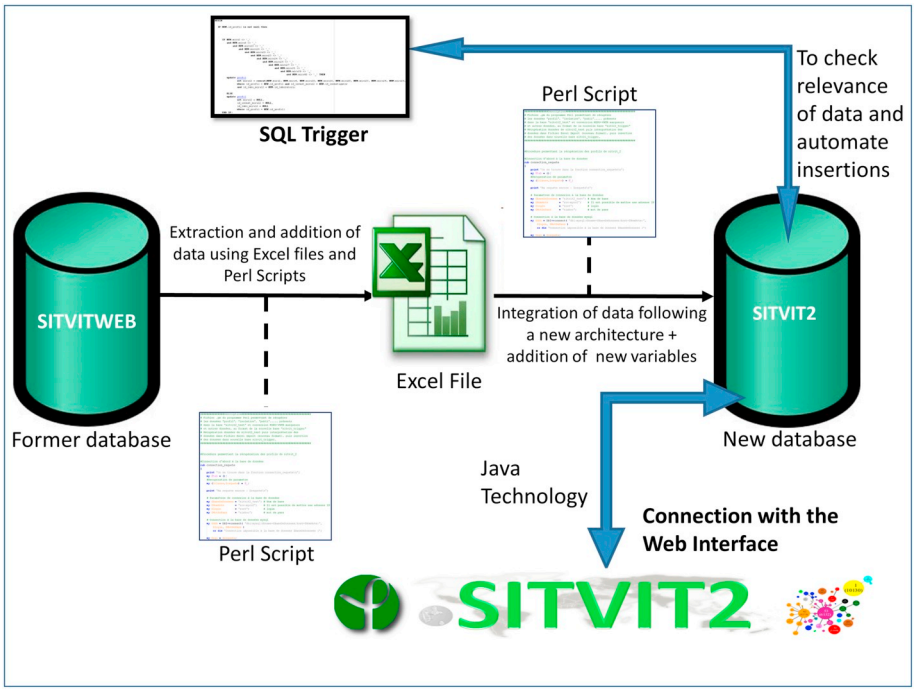

Fig. 1. Diagram representing the process of improvement of the new interface of our database, implementing a new MySQL architecture and adding new characteristics of patients.

epidemiological parameters were performed using Pearson's Chi-square test and Fisher's Exact Test. Student's *t*-test was used to compare continuous variables such as the age of patients. A binary logistic regression model was used to see which patient characteristics were potentially associated to drug resistant TB; drug resistance variable (meaning pansusceptible or drug-resistant TB) was used as dependent variable; and sex, age, lineage, and HIV serology were used as independent variables. Odds Ratios (OR) and 95% Confidence Intervals (CI) were calculated to assess the significance of the statistical tests. *P* values of < 0.05 were considered as statistically significant. Mean and standard deviation (SD) were also calculated for age of the patient.

Minimum Spanning Trees (MSTs) were constructed based on MIRU-VNTRs and spoligotyping patterns using BioNumerics software version 6.6 (Applied Maths, Sint-Martens-Latem, Belgium) and MLVA Compare (Ridom GmbH, Münster, Germany / GenoScreen, Lille, France). MSTs are connected undirected graphs in which all of the patterns are linked together with the fewest linkages between nearest neighbors. SpolTools software was used to draw Spoligoforests trees based on a Hierarchical Layout or the Fruchterman-Reingold algorithm (Reyes et al., 2008; Tang et al., 2008). Contrary to the MSTs, Spoligoforests are directed and not necessarily connected graphs, and permit to highlight the evolutionary relationships between ascendant and descendant spoligotyping patterns.

2.4. Web development and online access of the database

SITVIT2 was developed using the MySQL relational database management system (RDBMS). Perl language was employed to improve database entry, adding relevant information fields (demographic, drug resistance, HIV serology) from condensed data extracted from Excel files (Fig. 1). Tools and n-tier architecture of the Web application were implemented with Java technology or other web languages such as Java Server Pages (JSP), JavaScript, jQuery, Ajax, DHTMLX, and Google Application Programming Interface (API). JavaScript activation is recommended in order to launch SITVIT2 which brings new functionalities to interrogate the database for various added parameters. The graphical user interface (GUI) of SITVIT2 was deployed on the Apache Tomcat server (version 6), and the web pages were developed using Eclipse software (www.eclipse.org/). The SITVIT2 web application was successfully tested on Firefox and Google Chrome web browsers. SITVIT2 is freely accessible at: http://www.pasteur-guadeloupe.fr:8081/SITVIT2 (in case of problems to access this website, users are requested to verify with their network administrator to make sure that no restrictive firewalls were installed to block outbound connections to port 8081). If a troubleshooting is needed despite user guide provided (see Supplemental File 1), users may email the Web Content Manager.

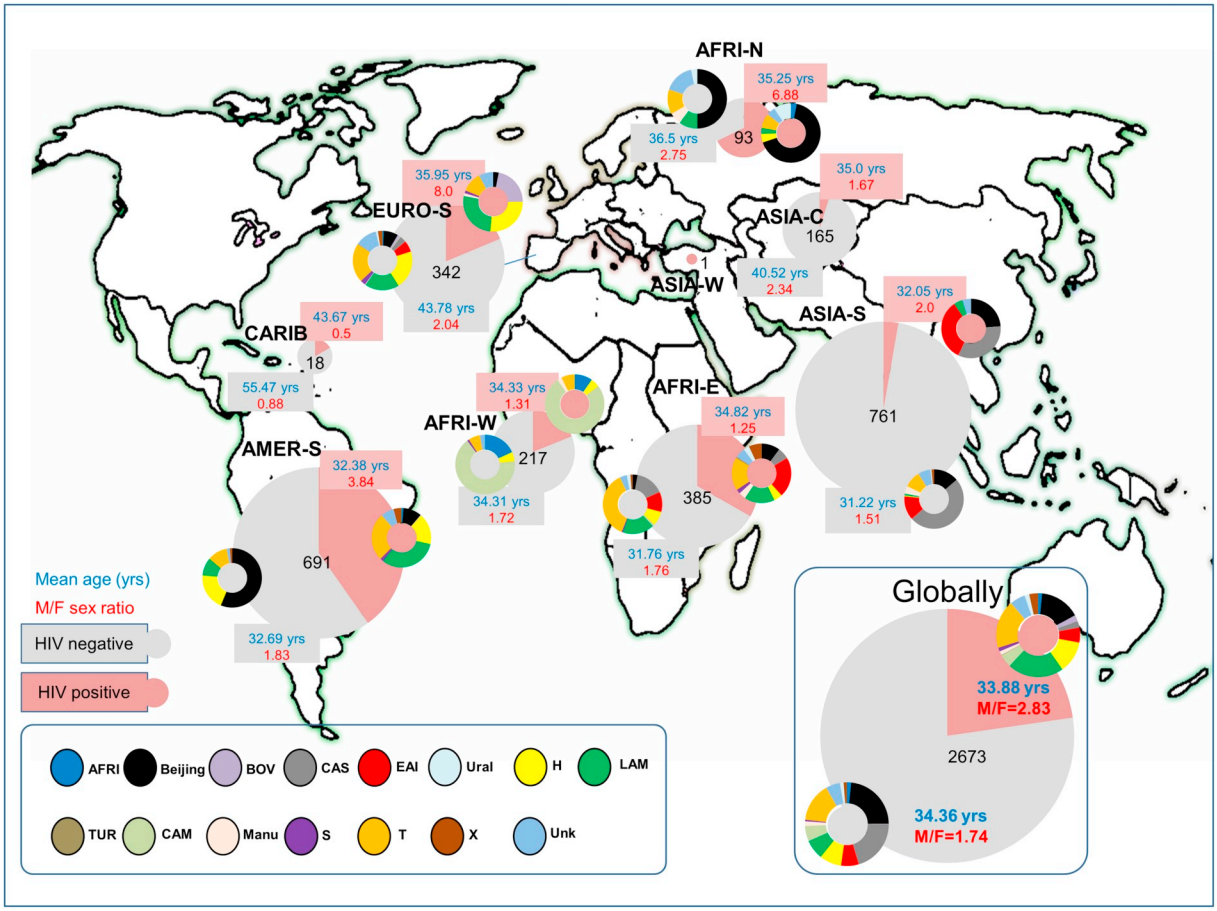

Fig. 2. Global Distribution of HIV-positive and negative patients in SITVIT2, based on spoligotyping phylogenetic lineages. The number within this pie chart shows the number of strains having information on HIV serology.

2.5. Drug resistance data

Drug susceptibility testing (DST) was performed using classical methods endorsed by the WHO, identifying first line and/or second-line anti-TB drugs. Recording of drug resistance information in SITVIT2 database was done as follows:

- DST code 1 designates pansusceptible strains (no drug-resistance).
- DST code 2 designates multidrug-resistant tuberculosis (MDR-TB), i.e., combined drug resistance to isoniazid (INH) and rifampin (RIF), with or without resistance to other drugs.
- DST code 3 designated resistance to any drug(s) excluding MDR or Extensively drug-resistant tuberculosis (XDR-TB, see below for definition).
- DST code 4 designates XDR-TB, defined as MDR-TB with additional resistance to any fluoroquinolone, and to at least one of the three following injectable drugs used in anti-TB treatment: capreomycin, kanamycin and amikacin.

Note that for analyses comparing drug-susceptible vs. drug-resistant strains, isolates corresponding to DST code 1 (pansusceptible) were compared against strains corresponding to DST codes 2 to 4 that were pooled collectively as drug-resistant isolates.

2.6. Geographical distribution

The worldwide distribution of *M. tuberculosis* complex strains was studied both country wise (2 or 3 letter country codes according to http://en.wikipedia.org/wiki/ISO_3166), and at macro-geographical level in subregions defined according to United Nations (available at http://unstats.un.org/unsd/methods/m49/m49regin.htm) as follows; Regions: AFRI (Africa), AMER (Americas), ASIA (Asia), EURO (Europe), and OCE (Oceania), subdivided in: E (Eastern), M (Middle), C (Central), N (Northern), S (Southern), SE (South-Eastern), and W (Western). In this classification scheme, CARIB (Caribbean or West Indies) belongs to Americas, while Oceania is subdivided in 4 sub-regions, AUST (Australasia), MEL (Melanesia), MIC (Micronesia), and POLY (Polynesia). Note that Russia was attributed a new sub-region by itself (Northern Asia) instead of including it among the rest of Eastern Europe.

3. Results and Discussion

3.1. HIV serology and associated parameters

The HIV serology was known for 2673 patients: 607 (22.71%) being HIV-positive and 2066 (77.29%) being HIV-negative. Isolates from HIV-positive cases were predominantly reported from Russia, South America and East Africa (at respective proportions of 67.74%, 40.23% and 33.25%). Among 16.67% to 18.89% strains submitted to SITVIT2 from West Indian, Southern European and West African populations were isolated from HIV-positive patients as compared to < 5% strains submitted from the Southern (Indian subcontinent) and Central Asia. HIV serology was not available for remaining subregions. Mean age of both HIV-negative and HIV-positive groups of patients was similar, representing respectively 34.36 yrs. (with a standard deviation of 15.62) and 33.88 yrs. (with a standard deviation of 8.69); nevertheless, age of patients varied in considerable proportions depending on the region of isolation (Fig. 2). For data submitted to SITVIT2, HIV-positive patients between 0 and 20 yrs. were essentially reported in Peruvian, Spanish, Mozambican and Nigerian populations. Nevertheless, the sex-ratio significantly differed between HIV-negative (1279/733 = 1.74 male/female sex ratio) and HIV-positive (438/155 = 2.83 male/female sex ratio) cases, with a p-value < .0001 (OR = 1.62; 95% Confidence Interval [1.32; 1.99]).

Regarding the whole dataset, when we compared the distribution of lineages among HIV-negative and HIV-positive patients (taking into account the ubiquitous T-family to make the comparison), we noticed that the LAM, X, and BOV lineages were significantly found at higher proportions among HIV-positive patients, with Odds ratios of 2.29 (95%CI [1.65; 3.18], p-value < .0001), 2.20 (95%CI [1.13; 4.24], p-value < .02), and 10.16 (95%CI [3.15; 42.97], p-value < .0001) respectively. On the other hand, Beijing and CAS lineages were not significantly associated with HIV-positive patients globally (p-value < .0002); one must however note a recent data entry of nearly 200 Beijing strains from a Peruvian study (Iwamoto et al., 2012) which could explain the high proportion of HIV-negative patients among Beijing isolates in South America. Conversely, LAM lineage was significantly associated with HIV-positivity in this area: representing 32.13% of all HIV-positive isolates vs. 9.80% of all HIV-negative isolates; X and S lineages were also more often isolated from HIV-positive patients.

The male/female sex-ratio among HIV-positive and HIV-negative patients of East Africa (1.25 vs. 1.76) did not differ significantly from that recorded in West Africa (1.31 vs. 1.72). Furthermore, unlike in South America, the proportion of LAM lineage was not significantly different among HIV-positive versus HIV-negative patients in East-Africa (16.41% vs. 17.12% respectively). T, CAS, and Haarlem lineages were more often present in isolates from HIV-negative patients, whereas EAI, Beijing, X, Manu, and S lineages were more frequent in strains from HIV-positive patients. As opposed to the AFRI lineage in West Africa, the Cameroon lineage (previously labeled LAM10-CAM) was more frequently isolated from HIV-positive patients (75.61% vs. 66.48%). LAM and BOV lineages were more commonly found among HIV-positive patients in Southern Europe; while Beijing and Ural lineages prevailed among HIV-positive patients of Russia. Although isolates submitted from Southern Asia corresponded to a low HIV-positivity rate, the proportion of EAI and Beijing lineages was relatively higher among HIV-positive patients. A single case of HIV-positive female patient from Western Asia (a 39 yrs. old patient from Turkey) corresponded to the ubiquitous SIT53/T1 genotype. Thus, the adaptability of the pathogen to the host could depend both on specific lineages but also to the geographic variations in host immunity. Nonetheless, utmost caution is needed before drawing definitive conclusions because of the lack of available data for all subregions (see the website for detailed comparisons).

Last but not least, a study of correlation between HIV serology and drug resistance (see subsection 3.2 below), did not show any definite correlation between HIV serology and drug resistance. Although this association may vary depending on the regions studied, data submitted to SITVIT2 suggests that drug resistant TB is not significantly associated with HIV-positivity at a global scale.

3.2. Drug resistance, genotypic lineages and associated parameters

Among the 19,606 available isolates containing information on drug resistance, we considered a sub-sample of 16,464 isolates from 142 countries of origin of patients for a deep analysis for which associated demographic and/or epidemiologic data were available (individualized analyses are possible by questioning the SITVIT2 web interface directly by users).

Regarding geographical distribution of drug resistance, of 16,464 isolates, a total of 5640 isolates were drug resistant (pooled DST codes 2 to 4). Among these, 2105 (37.32%) corresponded to DST code 3 (any drug resistance excluding MDR/XDR-TB); 3353 (59.45%) corresponded to DST code 2/MDR-TB, and 182 (3.23%) to DST code 4/XDR-TB. In our database, MDR strains were predominantly from countries that specifically undertook studies on characterization of MDR-TB, and hence not fully representative of the real snapshot of prevailing drug-resistance worldwide. DST code 2/MDR-TB isolates corresponded to 558 (62.42%) from India, 356 (29.67%) from Spain, 182 (42.42%) from China, 181 (47.01%) from Mexico, 151 (56.77%) from Russia, 147 (34.11%) from Turkey, 59 (80.82%) of strains from Kazakhstan, 55 (41.98%) from Venezuela, 39 (33.33%) from Poland, 37 (43.02%) from Dominican Republic, 18 (20%) from Guyana, and 16 (30.19%) from Kyrgyzstan. Furthermore, exclusively MDR-TB isolates were submitted to SITVIT2 from South Africa ($n = 785$), Ethiopia ($n = 74$) and Panama ($n = 37$). In this context, a recent study showed that MDR-TB in Panama is driven by clonal expansion and ongoing transmission of several strains of the LAM family, in which a specific strain (LAM9-c1) was shown to be closely related to an XDR strain identified in KwaZulu-Natal (KZN), South Africa (Lanzas et al., 2013). Lastly, in the remaining countries of isolation, the proportion of MDR-TB strains was lower than 20%, or not significant due to a smaller number of isolates (< 30 strains; Supplemental File 2).

We also looked if certain genotypes/lineages were more frequently concerned by drug resistance. Spoligotyping based MTBC lineages were used to underline certain geo-specificities with regard to the ability of specific lineages of being more resistant to anti-TB drugs than other lineages (Fig. 3). Globally, CAS, Beijing, Turkey, Bovis, and Manu lineages were more likely to be drug resistant (DST codes 2–4) to anti-TB drugs; while AFRI, Haarlem, T, and Ural lineages were associated with pansusceptible MTBC (p-value < .0001). Conversely, the remaining X, S, EAI, LAM, and Cameroon (formerly LAM10-CAM) lineage proportions were almost similar in both Pansusceptible and drug resistant groups. Nonetheless, if one focuses on individual subregions, a same lineage can act differently depending on the geographical region of isolation, e.g.:

(i) LAM and Haarlem (which are usually related to pansusceptible-TB in other regions), were found to be significantly associated to drug resistant-TB strains in the Caribbean;
(ii) As recently shown for Beijing lineage (Couvin and Rastogi, 2015), the association of Beijing isolates to drug resistance varied significantly depending on the region of isolation.
(iii) As opposed to the AFRI lineage, the Cameroon lineage was linked with drug resistant-TB in West and Central Africa;
(iv) EAI lineage was significantly associated to drug resistant-TB in North Africa, whereas this was not the case in Western and Southern Asia, the Caribbean, South America, and Western Europe;
(v) CAS lineage which could be the ancestor of Beijing family (Rastogi and Sola, 2007), was significantly associated with drug resistant-TB in Southern Asia, North Africa, and East Africa (specially Ethiopia); but not in Western Asia and Northern Europe.
(vi) Although isolates belonging to the S lineage constituted a very low proportion of strains in present study, we noticed that these isolates were more frequently associated to drug-resistance in the Caribbean and in Central America.

MSTs were drawn in order to highlight which patterns/lineages were more frequently concerned by drug resistance (DST codes 2 to 4) in SITVIT2 (note that supplemental MST Figures are accessible directly on the web tool).

Concerning gender and age of patients with drug resistant strains, the difference between the mean age of the patients infected with pansusceptible strains (44.87 yrs.; 95% Confidence Interval [44.29; 45.46]) vs. drug resistant strains (37.12 yrs.; 95% Confidence Interval [36.55; 37.68]) was significant regarding the whole study. Thus, TB patients with drug resistant strains were younger than TB patients with pansusceptible strains (p-value < .0001). However, this trend in mean age varied according to subregions; e.g., the difference between mean age of pansusceptible (P) versus drug-resistant (D) TB patients was significant in AFRI-N, ASIA-E, ASIA-S, ASIA-W, EURO-N, AMER-C (with p-values in the range of < 0.0001 to 0.026), but not in AFRI-W, AFRI-E, ASIA-W, EURO-S, EURO-W, CARIB, and AMER-S subregions. Furthermore, although the male/female sex-ratio was almost similar among pansusceptible and drug resistant cases at a global scale (1.84 vs. 1.87 respectively), significant variations were observed depending on the subregions studied (Fig. 3).

Considering HIV serology and drug resistance, results on HIV serology and drug resistance should be interpreted very cautiously since combined information on drug resistance and HIV serology was available for a limited number of countries/regions. In our sample, HIV-

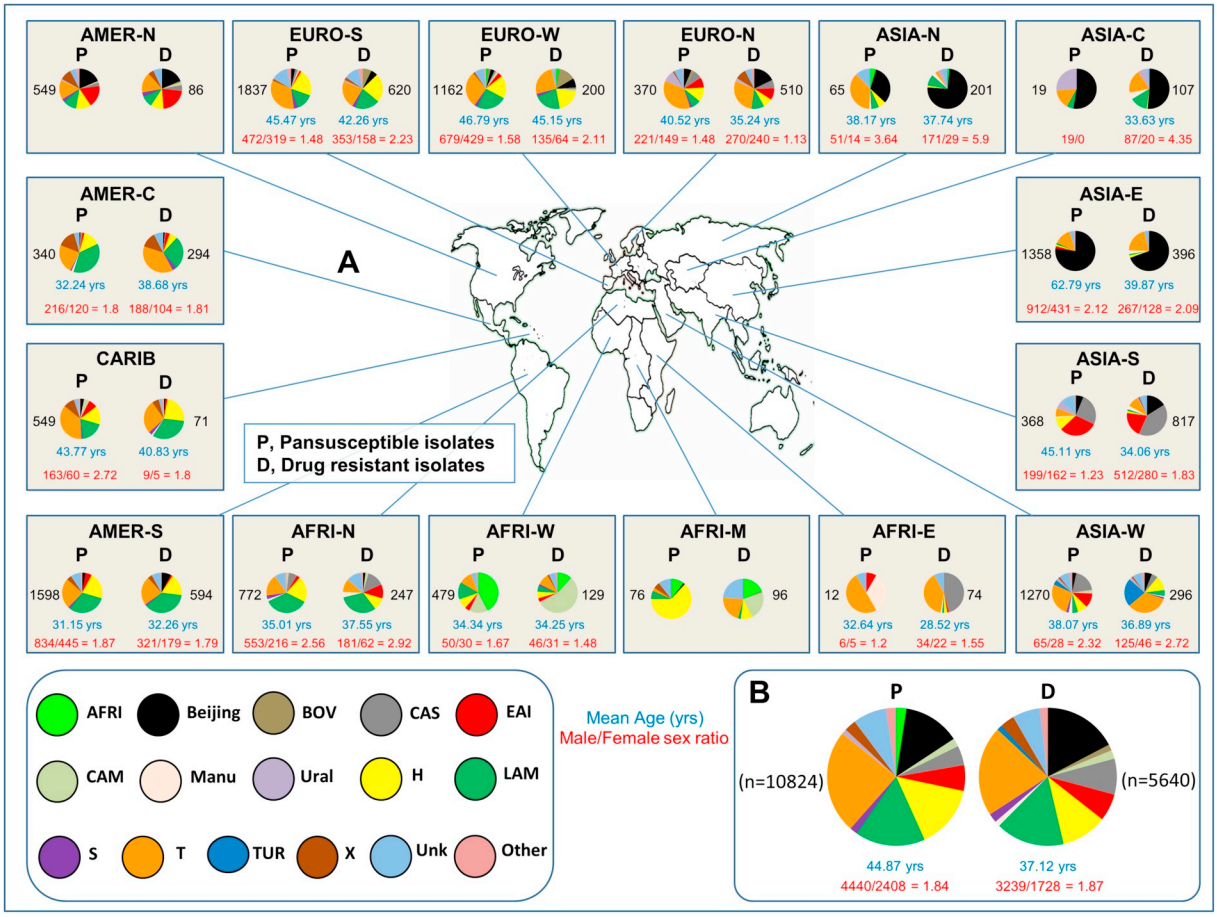

Fig. 3. Distribution of phylogenetic lineage and other parameters of the patients (age and sex) based on drug resistant (denoted by letter "D", corresponding to DST codes 2 to 4) vs. pan-susceptible strains (denoted by letter "P", corresponding to DST code 1) according to the UN sub-regions (A), and at the worldwide level (B).

positive serology was globally more often recorded for pansusceptible strains (23.21% vs. 8.65% within drug resistant group, p-value < .0001; OR 3.19, 95%Confidence Interval [2.15; 4.73]). At the level of subregions (number of HIV-positive/HIV-negative patients: East Africa, 7/0; Northern Asia, 57/30; South America, 0/3; Southern Asia, 14/622; West Africa, 30/125; and West Indies, 3/15); the difference in the proportion of HIV+ or HIV- serology was similar in pansusceptible or drug resistant-TB in West Africa and Northern Asia, HIV+ serology was linked to drug-resistant strains in the Caribbean, and HIV- serology among pansusceptible strains in Southern Asia.

Since SITVIT2 provides information on main MTBC lineages and the geographic origin of the patients, we also attempted to trace the potential pathways of transmission of drug resistant *M. tuberculosis* strains. Based on data available for 12,243 isolates with origin of patients and 4035 strains with drug resistance (DST codes 2, 3, or 4), Fig. 4 highlights the main subregions of origin of patients having a high proportion of drug resistant isolates, and potential migratory pathways as deducted from international migration flows (Abel and Sander, 2014) and United Nation's international migration report 2017 (https://www.un.org/development/desa/publications/international-migration-report-2017.html). In this figure, the thickness of the arrows roughly indicates the proportion of drug resistant strains, suggesting that a high proportion of patients carrying drug-resistant strains may came mainly from East–, North–, and West Africa; Southeastern–, Southern–, and Western Asia; Eastern Europe; and South America; and in direction of Northern–, Western–, and Southern Europe, and North America. Readers may consult the web version of the database for an overview on the distribution of drug-resistant TB by country, as well as in function of the lineages.

3.3. Genotyping data, lineages, demographics and phylogeny

3.3.1. Genotyping data and potential cleavages

SITVIT2 provides data recorded in 131 countries of isolation (containing 1032 cities of isolation) derived from 169 countries of origin of patients, as follows: spoligotyping for 103,856 isolates (3851 SITs containing 98,049 isolates +5807 orphans); 5-locus ETRs for 10,294 isolates (424 VITs containing 9539 isolates +755 orphans); 12-loci MIRU-VNTRs for 19,311 isolates (1632 12-MITs containing 14,777 isolates +4534 orphan isolates); 15-loci MIRU-VNTRs for 4702 isolates (430 15-MITs containing 1827 isolates +2875 orphans); and 24-loci MIRU-VNTRs for 1672 isolates (178 24-MITs containing 765 isolates +907 orphans) (Table 1). Note that strains with incomplete MIRU-VNTRs loci information were considered as orphan in the database. Detailed information about the genotyping markers and the worldwide distribution of specific genotypes, and lineages/sub-lineages is accessible by interrogating the online database. Furthermore, regarding correspondence with MIRU-VNTRplus nomenclature, all our 15- and 24-loci MIRU-VNTRs (having a MIT number were entered in MIRU-VNTRplus database, and were assigned an identifier based on MtbC15–9 Nomenclature. The files describing these MIRU-VNTRs patterns are available online (http://www.pasteur-guadeloupe.fr:8081/SITVIT2/tools.jsp). Lastly, the SITVIT2 data was linked to SITVIT-KBBN lineages using the SITVIT-KBBN tool (Aminian et al., 2014), explicitly developed to improve our expert-based classification of the genotypes, using a Bayesian network. This project was conjointly performed by Institut Pasteur de la Guadeloupe and Rensselaer Polytechnic Institute (RPI), and is consultable at: http://tbinsight.cs.rpi.edu/run_tb_lineage.html (select the option "SITVIT Clade by KBBN").

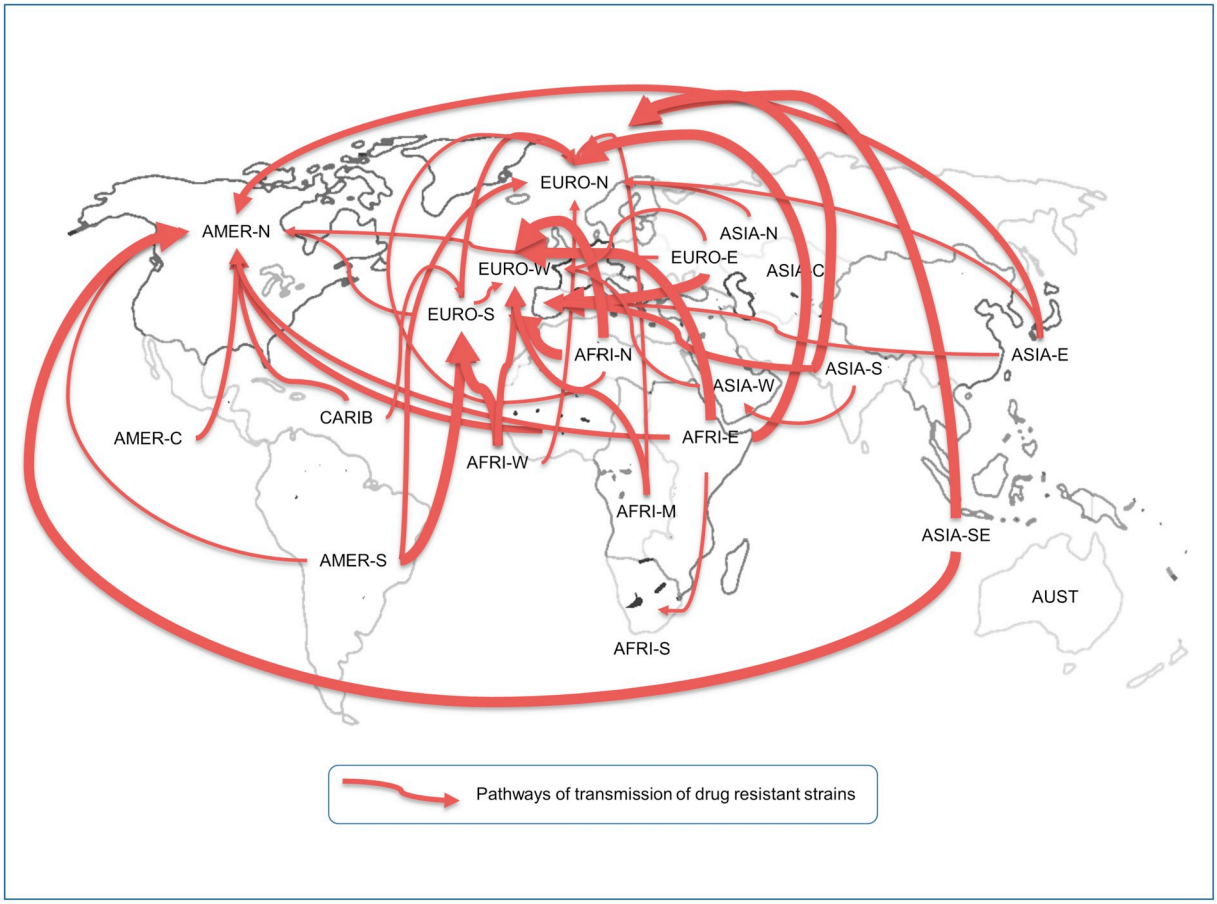

Fig. 4. Map representing the potential main migrations of drug resistant-TB strains. Thickness of the arrows roughly indicates the number of drug resistant strains originating from one region of origin to the region of isolation. Number of drug resistant strains between 5 and 10, between 10 and 20, and > 20 were represented by thin, medium, and large arrows respectively.

Supplemental File 3 provides a global snapshot of geographical cleavages that may exist between various geographical areas in function of sex-ratio and age-groups of the patients. Age range of 0–20 yrs. was predominantly found among patients from Latin America (Central and South America) and Indian sub-continent (reaching almost 20% of patients); age range of 21–40 yrs. was predominant (around 60%) in Africa (West, North and East Africa); while patients were older in Eastern Asia (especially Japan), and European countries, as opposed to African countries. Some lineages/sub-lineages were more common among younger patients (0–20 yrs) at the macro-geographical level, e.g., BOV_1 (22.22%), CAS1-Delhi (18.53%), CAS1-Kili (20.00%), LAM3 (16.58%), LAM6 (15.32%), Turkey or former LAM7-TUR (14.02%), T2-Uganda (17.65%), T3-ETH (18.92%), T5-Madrid2 (24.24%), LAM-RUS (14.04%), Ural-1 (14.88%), X1 (15.70%), and X3 (15.87%). Regarding the mean age characteristics, significantly lower mean age of 32.05 yrs. was observed for patients with CAS1-Kili lineage, followed by 32.77 yrs. for AFRI_1 lineage. Lastly, despite an overall higher proportion of male than female patients globally, some lineages/sublineages were well represented among female patients. As regards to the global sex-ratio distribution in function of lineages, the male/female (M/F) sex-ratio was significantly lower for T3-ETH, Ural-2, Manu1, AFRI_2, EAI2-Manila, BOV_1, T4, and LAM4, with M/F sex-ratios between 1.03 (38/37) to 1.39 (152/109). Note that detailed information is accessible by querying the database online.

The MST in Supplemental File 4 shows all spoligotypes recorded in SITVIT2, according to their phylogenetic lineages (color nodes) and their distribution in the various subregions (separations inside nodes). We note that all major lineages were relatively well distinguishable. The patterns belonging to the "Euro-American" lineage (T, S, LAM, X, Haarlem, Ural, Cameroon, Turkey) were found primarily in the upper and central parts of the tree, while some outlying clusters of the tree concerned isolates of ancestral lineages (AFRI, BOVIS, CAS, EAI). Groups of spoligotypes belonging to AFRI and BOVIS lineages were particularly close. We also noticed that patterns carrying "unknown" lineage signatures were often positioned adjacent to the "Euro-American" spoligotypes; while a smaller yet significant number of strains was placed close to the Manu family. The spoligotypes belonging to the Cameroon lineage (formerly known as LAM10-CAM) were closer to the genotypes of the T lineage. The MST constructed in Supplemental File 5 shows genetic diversity of all 12-loci MIRU patterns recorded in SITVIT2. This figure highlights predominant 12-MITs which are well visible in the tree as "central nodes" and identify main 12-loci phylogenetical lineages described in a recent work (Hill et al., 2012). Recently updated Ethiopian lineage (see subsection 3.4 below) was not visible in the MST trees. Lastly, Table 2 briefly indicates the predominant 12-MITs in our database (containing 50 or more isolates), associated spoligotyping based major lineages in SITVIT2, and corresponding lineages attributed by the MIRU-VNTRplus web tool (Allix-Béguec et al., 2008; Weniger et al., 2010). Despite some rare divergences (highlighted in gray in Table 2), lineages attributed by SITVIT2 matched well with those attributed by MIRU-VNTRplus. Thus additional comparisons could be performed to link the data between the two databases thanks to SITVIT2.

3.3.2. *Trends over time and potential movements of strains*

Comparison of lineage distribution in SITVITWEB vs. SITVIT2 may underline the movement of certain patient populations carrying a specific lineages. For example, one can observe a relative progression of the recently defined Turkey lineage into areas surrounding Turkey in

Table 2
Main 12-MITs (containing at least 50 isolates) and their associated lineages based on SITVIT2 and MIRU-VNTRplus databases.

12-MIT	MIRU-VNTRs	Nb in DB	Predominant Spoligotyping based Lineages in SITVIT2 (%)*	Lineage in MIRU-VNTRplus
17	223325173533	479	Beijing (97.09)	Beijing
45	225325153323	447	H3 (44.6), H1 (37.77)	Haarlem
16	223325153533	331	Beijing (97.61)	Beijing
33	224325153323	272	X2 (54.37), T1 (17.5), H3 (12.5)	?
68	225425173533	226	CAS1-Delhi (52.94), CAS (32.35)	?
8	223125153324	221	T1 (74.36)	?
42	225313153323	217	H3 (63.76), H1 (13.42), H2 (11.41)	Haarlem
43	225323153323	207	H1 (52.24), H3 (28.36), H2 (5.97)	Haarlem
34	224325153324	173	X1 (32.26), X3 (26.88), X2 (15.05), T1 (11.83)	?
40	225125113322	155	T1 (65.71), T3-ETH (17.14), T2 (7.14)	TUR
56	254326223432	154	EAI2-Manila (69.15), EAI2-nonthaburi (14.89)	EAI
25	224226153321	147	LAM9 (54.33), LAM5 (23.62), LAM4 (5.51)	LAM
213	224326153323	144	LAM3 (86.52)	LAM
190	124326153220	106	LAM9 (95.19)	LAM
144	224325153314	103	X1 (50.0), T1 (31.82)	?
163	224126152321	100	LAM11-ZWE (62.5), LAM9 (15.63)	LAM
246	124326153324	96	LAM4 (86.96), LAM9 (6.52)	LAM
157	223425153322	95	T1 (70.0), T3 (8.0), H3 (8.0)	?
261	227425113434	95	CAS1-Kili (90.0), CAS (10.0)	?
104	222325173543	88	Beijing (100.0)	Beijing
224	223326153321	86	T1 (85.71), T2 (5.71)	LAM
128	223226153321	85	LAM1 (67.44), LAM9 (18.6)	LAM
271	226425153533	84	CAS1-Delhi (86.36), EAI5 (9.09)	?
46	225325153324	83	H1 (46.0), H3 (22.0), X3 (14.0), T1 (8.0)	Haarlem
7	222325153323	82	Ural-2 (22.0), T1 (20.0), H3 (18.0), H1 (16.0)	?
69	254326223434	82	EAI3-IND (100.0)	EAI
318	226425173533	81	CAS1-Delhi (76.92), CAS2 (7.69), CAS (7.69)	Delhi/CAS
117	224325143324	80	T4-CEU1 (47.06), X2 (17.65), T1 (11.76), X1 (9.8)	?
15	223325153322	79	T1 (72.73), T2 (7.27)	?
310	215125113322	78	Tukey (57.97), T1 (21.74), T3-OSA (18.84)	TUR
152	225325153322	77	H3 (45.16), H1 (32.26), T1 (6.45)	Haarlem
194	214125113322	77	Turkey (92.11)	TUR
633	224315153324	76	T2 (89.19)	Cameroon
156	223125153322	74	T1 (44.0), T5 (32.0), T3 (8.0), T2 (8.0)	?
374	226425173423	73	CAS1-Delhi (83.33)	Delhi/CAS
83	223325163533	72	Beijing (100.0)	Beijing
32	224325153322	71	X1 (38.1), X2 (16.67), T1 (9.52), X3 (7.14)	Haarlem
12	223315153323	69	Cameroon (89.19)	Cameroon
135	221325173533	68	Beijing (100.0)	Beijing
140	124326153224	67	LAM9 (20.93), T1 (18.6), LAM-RUS (9.3), T-Tuscany (6.98), LAM6 (6.98)	LAM
30	224325143323	63	X2 (30.77), T1 (15.38), LAM (11.54), H1 (11.54), LAM5 (7.69)	?
1211	225425183324	61	H3 (93.44), H1 (6.56)	?
116	223325153323	60	T1 (31.43), X2 (20.0),H3 (14.29), X1 (11.43), H1 (5.71), T2 (5.71)	?
64	254326223513	57	EAI1-SOM (59.26)	EAI
26	224226163321	54	LAM2 (51.85), LAM1 (27.78), LAM9 (11.11)	LAM
112	223325143324	53	X1 (36.85), T3 (21.05), T1 (18.42), T4 (5.26), T2 (5.26)	?
237	223126152321	53	LAM11-ZWE (36.36), LAM4 (36.36), LAM9 (27.27)	LAM
396	223125153323	53	T1 (52.17), LAM3 (43.48)	?
456	224315153321	53	Cameroon (66.67), T2 (33.33)	Cameroon
212	233325153324	51	S (82.5), LAM3 (5.0)	S
35	224325153325	50	H3 (27.78), Ural-2 (22.22), T1 (11.11), X1 (11.11), X3 (8.33), LAM9 (5.56), T2 (5.56)	X
231	223125143324	50	T1 (71.43), T (14.29), H2 (7.14)	?

*Differences between lineages attributed by SITVIT2 as compared to MIRU-VNTRplus are highlighted in gray.

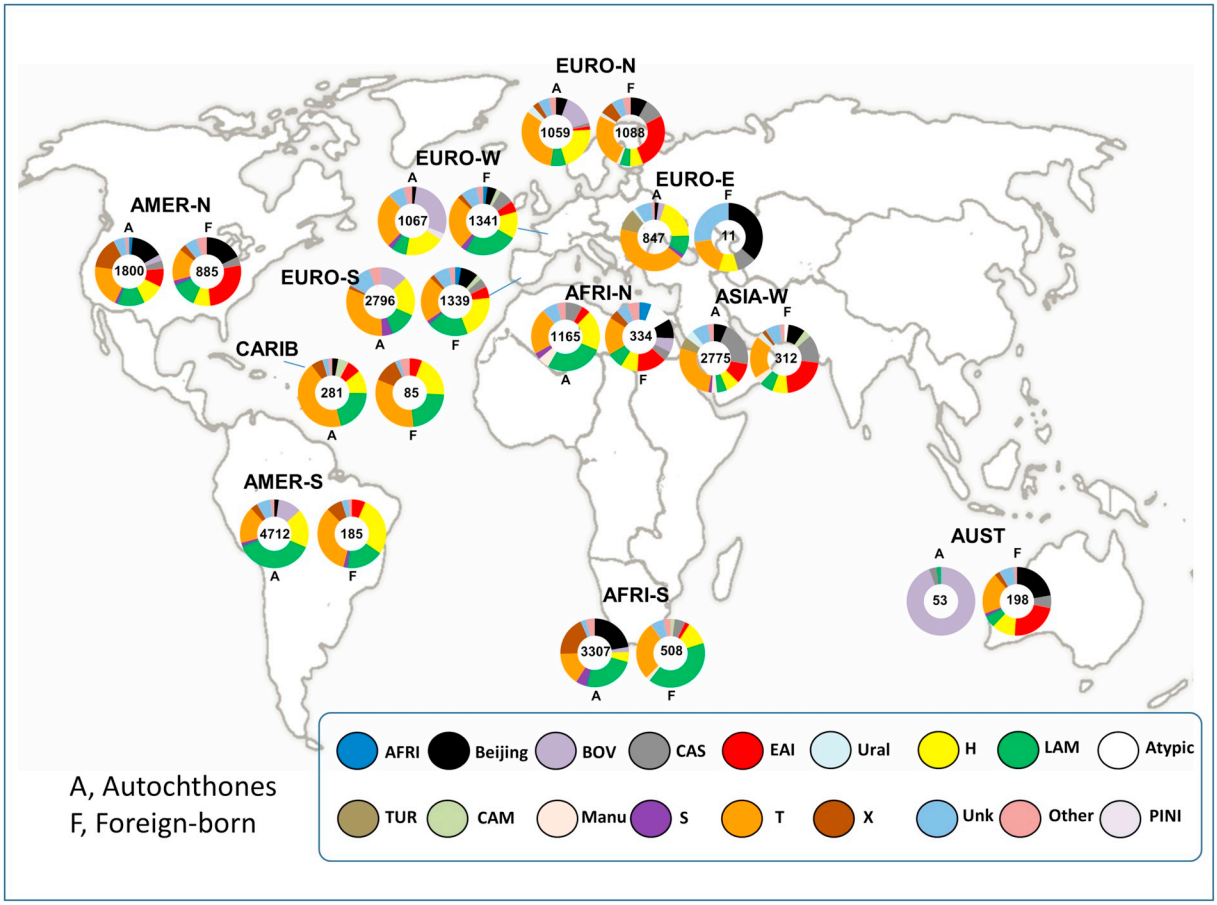

Fig. 5. Phylogeographical distribution of lineages in function of origin of patients (autochthones vs. foreign-born). The number included inside rings represents the number of strains involved in each group of patients. Lineages representing > 1.5% of isolates in a given sub-region were taken into account.

Western Asia (such as Eastern Europe). Drifts of evolution of lineages or groups of lineages over time may give us an overview of trends over time and space, similar to what was shown recently regarding the worldwide increase of MDR/XDR strains associated to the Beijing lineage (Couvin and Rastogi, 2015). Similar trends of evolution of lineages in function of drug resistance and/or over time, can be visualized on our web page (http://www.pasteur-guadeloupe.fr:8081/SITVIT2/stata.jsp). Another tool allows users to get a global overview on what is happening through time and space concerning the lineages contained in SITVIT2, as well as the evolution of isolates in function of drug resistance information, sex or age groups of patients. Thus data discernable by navigating on thumbnail "Other Statistics" (on the SITVIT2 website) allows to know which lineages might be progressing in specific regions. By observing evolution of strains in function of lineages, age, sex and associated demographic characteristics in different geographic areas, one can easily estimate if TB incidence is progressing, and identify the priority subpopulations concerned (gender, age groups, origin, etc.).

Information on the origin of patients is frequently available in SITVIT2 which allows to have an insight on the impact of migration of population on worldwide TB distribution and population structure. For example, the distribution of lineages of autochthones vs. foreign born patients (Fig. 5) provides with an overview of the global imported cases of TB (mainly in regions/countries in which TB has a low incidence). As underlined earlier, the persistence of the TB epidemic in most of the low incidence countries today is largely attributable to imported cases from patients originating from high TB incidence areas (García de Viedma et al., 2011). One may notice that the EAI lineage was often observed from foreign-born patients in Northern, Western and Southern Europe, as well as in North Africa, Western Asia, and North America. One may also note the presence of strains belonging to CAS lineage among foreign-born patients in Northern, Western, and Eastern Europe. Haarlem and X lineages were the most visible lineages prevailing among foreign-born patients from the Caribbean, South America and Austral Africa. LAM lineage was widely spread among autochthonous patients from South America, but linked to foreign-born patients in Austral Africa; the latter observation underlines migration of people from countries such as Zimbabwe within Africa, where LAM lineage predominates (Zimbabwe is characteristic for having a high proportion of strains belonging to LAM11-ZWE sub-lineage). Lastly, a noticeable presence of strains belonging to BOVIS lineage was visible among autochthonous patients from Northern, Western and Southern Europe, as well as among patients from South America and Australasia (Fig. 5).

Knowing the patient origin of *M. tuberculosis* genotypes may also help to deduce history underlying the impact of migration, wars and colonization on prevailing population structure. One can take the example of controversial SIT4 ("undefined" lineage in SITVITWEB) mostly found in Southern Europe (predominantly in Italy; $n = 33$), to delineate a potential pathway of its spreading through a cycle of displacement of populations. In the supplemented SITVIT2 version, this genotype is found in Western Asia (particularly in Turkey, $n = 29$), East and Austral Africa, South and North America, as well as the rest of Europe (Western, Eastern and Northern), but not in Central America, and only sporadically in the Caribbean and the rest of Africa and Asia. This distribution may underline a cycle of displacement of populations animated by successive waves of exploration, trade, or colonization, such as Italian colonialism in Africa of the present day countries of Libya, Ethiopia, Eritrea, and Somalia, that lasted from 1890 to 1941, or even earlier during the height of Roman Empire. Furthermore, the high rate of SIT4 genotype observed in Turkey is probably due to the

Table 3
Reclassification of some spoligotyping lineage in SITVIT2 as compared to SITVITWEB.

SIT[a]	SITVITWEB	Former classification (In-house SITVIT2)	Correction (Final SITVIT2)
125	T2/LAM3 (controversial signature)	T2/LAM3 (controversial signature)	S
127	H4	Ural-2	Ural-2 (NEW-1)
159	T1	Unknown	T-Tuscany
254	T5-RUS1	T5-RUS1	LAM-RUS
266	T1	T1	LAM-RUS
343	Unknown	Unknown	Ethiopian
910	Unknown	Unknown	Ethiopian
1729	Unknown	Unknown	Ethiopian
3099	Not defined	EAI5	LAM9
3336	Not defined	Unknown	Ethiopian
3364	Not defined	Unknown	Ethiopian

[a] Please note that SITs 3099, 3336, and 3364 were missing in SITVITWEB.

relationships existing between the two countries since the times of the Ottoman Empire. Nonetheless, this hypothetical scheme is yet based on spoligotyping alone, and should be corroborated using MIRU-VNTRs and New Generation Sequencing (NGS).

3.4. Revised rules of prototype lineages, and other analyses

Thanks to the phylogenetical analyses and the visualization tools, we were able to redefine certain binary spoligotyping rules (Supplemental File 6). Furthermore, recent newly-defined sub-lineages such as Ural-1 and Ural-2 (Mokrousov, 2012), Turkey family (Abadia et al., 2010; Kisa et al., 2012), and Cameroon lineage (Koro Koro et al., 2013) were included. The designation of the EAI1-MYS lineage that was recently proposed to identify a genotypic family specific to Malaysia (Ismail et al., 2014), was added in the revised spoligotyping rules. These updated rules (Supplemental File 6), were subsequently implemented on all spoligotype patterns with a well defined lineage in SITVIT2. The study of the overall frequency, based on the schematic representations of spoligotypes as "SpoligoLogos" (Driscoll et al., 2002), and the presence or absence of different spacer according to the sublineages, allowed us to issue new hypotheses about the spacers generally present or absent. Black squares and red squares (filled or empty), designating the 43 spacers of spoligotypes, have the same meaning as in the previously described rules (Demay et al., 2012). However, blue squares (empty or filled) denote spacers that were respectively absent or present according to the lineages. These new rules could facilitate interpretation of spoligotypes and assignment to specific lineage. Nevertheless, we must admit that there are still unavoidable uncertainties in the interpretation of spoligotypes with long deletions of spacers (Mokrousov et al., 2014). We are aware that spoligotyping based lineages may be imperfect. Some changes between SITVITWEB and SITVIT2 lineage assignation are visible in Table 3 (a regularly updated list of spoligotypes will be available on the website; changes will be highlighted in red). An exhaustive list of differences that occurred between SITVITWEB and SITVIT2 (regarding spoligotyping lineage assignation) is reported in Supplemental File 7. In this new version (SITVIT2), corrections have been made according to recent suggestions by Mokrousov, 2018. We are also aware that the use of spoligotyping alone is not enough to accurately discriminate various TB strains (as mentioned earlier). In SITVIT databases, LAM9 sub-lineage can be considered as representing the general LAM lineage. Similarly, EAI5 sublineage can represent the general EAI lineage. SIT127 which was classified as belonging to H4 in SITVITWEB has then been reclassified as Ural-2 in SITVIT2. It could now be reclassified as Ural-2 (NEW-1). One may notice that H4 sub-lineage is no more visible in the updated SITVIT2 (Supplemental File 7). Further corrections will be applied to the new lineage named NEW-1. Spoligotyping patterns belonging to Ethiopian lineage (known as L7) (Firdessa et al., 2013) have been added to SITVIT2 according to recent research (Tulu and Ameni, 2018). The spoligotyping binary rule corresponding to this lineage has also been added in the updated rules (Supplemental File 6). Spoligotypes of this family could be easily searched from SpolSimilaritySearch using the pattern "■■■□□□□□□□□□□□□□□□□□□□□□■..□□.....■....." (i.e. presence of spacers 1–3, 25, 36, and absence of spacers 4–24, 28–29; dots could be replaced by dashes representing uncertainty on the web tool).

We can easily notice that certain genotypes were geographically more confined than others. This is why we attempted to develop an algorithm allowing the highlighting of rare identified genotypes (http://www.pasteur-guadeloupe.fr:8081/SITVIT2/tools.jsp). SIT1044/BOV_1, SIT566/T2, and SIT219/T1, are some examples of confined genotyping patterns in SITVIT2 database, they have been retrieved when looking for spoligotypes accounting for at least 50 isolates in our database and being present in at most 2 sub-regions. Some questions regarding rare genotypes remain unanswered. We can wonder, for example, why SIT14/X3 is predominant among isolates from Guadeloupe (Ferdinand et al., 2013; Millet et al., 2014). Moreover, SIT14 is exclusively found among strains isolated in the Americas. The spread of this particular profile could be due to the history of the Caribbean colonization, notably the conflicts that occurred during the 1700s, between French and English settlers. For example, in 1700, Guadeloupe belonged to French, whereas it belonged to Great Britain in 1759 (http://en.wikipedia.org/wiki/History_of_the_Caribbean). Further investigations could be necessary to assess whether this specific genotype is well represented in neighboring islands. Supplemental File 8 is a spoligoforest, drawn as a "Hierarchical Layout" (using SpolTools software), showing potential evolutionary scenarios of X lineage strains being available in SITVIT2 and having a SIT number (n = 4540 strains grouped into 147 SITs). When focusing on the Caribbean region, we have noted that certain SITs belonging to X3 sublineage appeared to be specific to 2 territories of the Caribbean: (i) SIT14 was predominantly found in Guadeloupe (n = 27/403 or 6.7%), (ii) SIT70 was predominantly found in the Bahamas (n = 8/28 or 28.57%). Pending complementary analysis, we suggest that (a) SIT14/X3 could be reclassified as SIT14/X3-GLP (characterized by the absence of spacer 2), and (b) SIT70/X3 as SIT70/X3-BHS (characterized by the absence of spacer 39). Interestingly, SIT13/X3-GLP (characterized by absence of spacers 2 and 15) is a descendant of SIT14 in the spoligoforest (Supplemental File 8). When querying all available X3 lineages strains containing both spoligotypes and 12-loci MIRU-VNTRs (http://www.pasteur-guadeloupe.fr:8081/SITVIT2/query), we noted that unlike other spoligotypes associated with X3 lineage group, SIT14 was rather linked with 12-MIT28 (value: 224325133325).These observations could suggest that there exist X3 lineage isolates that could be specific to the Caribbean region, e.g. SIT1366 – a descendant of SIT70 (Supplemental File 8) – is present in Panama (n = 4/179 or 2.23%). However, further investigations are obviously needed to endorse these observations. Note that many other similar queries are consultable by questioning the online web interface.

Of interest, the Haarlem lineage, firstly described in the Netherlands (Kremer et al., 1999) and highly prevailing in Europe, was significantly associated to drug resistance in the Caribbean, (and to a lesser extent in Western Asia). Haarlem is believed to have been introduced in Central Africa during European colonization (Filliol et al., 2003); in our study, this last lineage was globally prevalent among pansusceptible TB strains.

As underlined recently (Stucki and Gagneux, 2013; Kohl et al., 2014;van Soolingen, 2014; Coll et al., 2014b), multidisciplinary approaches and epidemiology based on multiple genetic markers and techniques are revolutionizing and becoming the new era of "gold standard" for studying evolutionary relationships, epidemiological links and transmission pathways of members of *M. tuberculosis* complex. Even though SNPs might be ideal for defining deep phylogenetic groupings, they offer insufficient discriminatory power for routine molecular epidemiological investigation in genetically monomorphic bacterial pathogens (Achtman, 2008; Comas et al., 2009). CRISPR-, VNTR-, and other related genotyping methods could thus likely remain important

genotyping tools for epidemiological purposes (García de Viedma et al., 2011; Jagielski et al., 2016). Therefore, the development of web accessible databases for global TB surveillance based on various classical/conventional and/or Whole Genome Sequencing (WGS) genotyping data remains highly relevant.

In the above context and despite its known limitations (see below), the SITVIT2 database accumulates a considerable amount of data on worldwide MTBC genotypic diversity in conjunction with available epidemiological and demographic information and drug resistance, thanks to the huge amount of information generated by scientists from all over the world. It exclusively includes data generated by PCR-based spoligotyping and MIRU-VNTR techniques which were previously considered the most easily applicable, simple and widely affordable techniques allowing study of large populations of samples (Blouin et al., 2012). Although the reliability of spoligotyping and MIRU-VNTRs typing techniques to identify circulating MTBC isolates has been underlined (Oelemann et al., 2007; Alonso-Rodriguez et al., 2009; Yuan et al., 2013), additional markers such as single nucleotide polymorphisms (SNPs) and Large sequence polymorphisms (LSPs) are helpful to exhaustively distinguish all patterns/lineages involved in TB transmission (Comas et al., 2009). Nevertheless, some limitations of our study could be the fact that MIRU-VNTRs patterns were not available for all strains, and the data collected is not truly representative of the global MTBC diversity and there may exist a potential bias in comparative analysis based on distribution of genotypes in function of: (i) HIV serology, (ii) drug resistance, and (iii) distribution of gender and age groups of patients. Nonetheless, a great effort was made to add data covering significantly more countries/regions as compared to the previous versions of the database. The establishment of a WHO-sponsored consortium of researchers working on MTBC genotyping around the world could be one of the ways to improve the representativeness of this database in the future.

In conclusion, even though the information available in SITVIT2 cannot be expected to provide robust enough phylogenetic support to classify each and single MTBC isolate, it does offer a high level of discrimination as well as correct and unbiased clustering assignment based on user's data. Data export as excel (XLS) file is possible through this web application. Thus despite generation of massive “omics” data in conjunction with efficient platforms associated with NGS techniques presently, classical methods of analysis should remain useful thanks to the SITVIT2 repository and web-tool. More significantly, SITVIT2 database will allow that huge amount of data generated using older genotyping methods are not simply deprecated but instead are useful to establish a more complete picture and knowledge on DNA based data by establishing useful links between the “older” and “newer” methodologies, and could be correlated to/with these new techniques, in order to provide a more complete picture and knowledge on DNA based data, thus facilitating a more comprehensive global overview on TB molecular epidemiology.

4. Conclusion

SITVIT2 database represents a huge collection of epidemiologic, demographic and clinical data on MTBC allowing an overview of worldwide distribution of *Mycobacterium tuberculosis* complex isolates. Our research is focused to improve in-depth phylogenetic characterization of MTBC lineages in conjunction with epidemiological analysis of circulating clones to generate evidence-based geographical mapping of predominant clinical isolates of tubercle bacilli causing the bulk of the disease both at country, regional and local level. Further superimposition of these maps with socio-political, economical, and demographical characteristics available through Geographic Information Systems (GIS) allows to have a precise view of prevailing disparities as seen at the level of United Nation's sub-regional stratification. An in-depth comprehension of these disparities and drawbacks is important to take appropriate actions by decision-makers and public health authorities alike, in order to better monitor, understand and control the tuberculosis epidemic worldwide.

Acknowledgements

We thank > 500 investigators who provided data for SpolDB4, SITVITWEB, and SITVIT2 databases (among which 190 contributed genotyping data on > 100 strains). Note that each strain with its genotyping information in these respective databases directly refers to the investigator in question (the full list is available upon request). We are highly grateful to Christophe Sola, Julie Millet, Véronique Hill, Elisabeth Streit, Yann Reynaud, Thibault Cousin, and Iñaki Cervera-Marzal for helpful discussions and collaboration. DC was awarded a Ph.D. fellowship by the European Social Funds through the Regional Council of Guadeloupe, while the work done at Institut Pasteur de la Guadeloupe was supported by a FEDER grant, financed by the European Union and Guadeloupe Region (Programme Opérationnel FEDER-Guadeloupe-Conseil Régional 2014-2020, Grant number 2015-FED-192). NR supervised and participated in the design of the study. DC and AD participated in the management of the architecture of the database, and designed the algorithms allowing to query the web application. TZ validated the spoligotyping rules. DC and NR wrote the present manuscript. All authors read and approved the final manuscript.

Conflict of interest

The authors declared that they have no competing interests.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.meegid.2018.12.030.

References

Abadia, E., Zhang, J., dos Vultos, T., Ritacco, V., Kremer, K., Aktas, E., Matsumoto, T., Refregier, G., van Soolingen, D., Gicquel, B., Sola, C., 2010. Resolving lineage assignation on *Mycobacterium tuberculosis* clinical isolates classified by spoligotyping with a new high-throughput 3R SNPs based method. Infect. Genet. Evol. 10, 1066–1074. https://doi.org/10.1016/j.meegid.2010.07.006.

Abel, G.J., Sander, N., 2014 Mar 28. 2014. Quantifying global international migration flows. Science 343 (6178), 1520–1522. https://doi.org/10.1126/science.1248676.

Achtman, M., 2008. Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. Annu. Rev. Microbiol. 62, 53–70. https://doi.org/10.1146/annurev.micro.62.081307.162832.

Allix-Béguec, C., Harmsen, D., Weniger, T., Supply, P., Niemann, S., 2008. Evaluation and strategy for use of MIRU-VNTRplus, a multifunctional database for online analysis of genotyping data and phylogenetic identification of *Mycobacterium tuberculosis* complex isolates. J. Clin. Microbiol. 46, 2692–2699. https://doi.org/10.1128/JCM.00540-08.

Alonso-Rodriguez, N., Martínez-Lirola, M., Sánchez, M.L., Herranz, M., Peñafiel, T., Bonillo Mdel, C., Gonzalez-Rivera, M., Martínez, J., Cabezas, T., Diez-García, L.F., Bouza, E., García de Viedma, D., 2009. Prospective universal application of mycobacterial interspersed repetitive-unit-variable-number tandem-repeat genotyping to characterize *Mycobacterium tuberculosis* isolates for fast identification of clustered and orphan cases. J. Clin. Microbiol. 47, 2026–2032. https://doi.org/10.1128/JCM.02308-08.

Aminian, M., Couvin, D., Shabbeer, A., Hadley, K., Vandenberg, S., Rastogi, N., Bennett, K.P., 2014. Predicting *Mycobacterium tuberculosis* complex Clades using Knowledge-based Bayesian Networks. Biomed. Res. Int. 2014 (398484). https://doi.org/10.1155/2014/398484.

Blouin, Y., Hauck, Y., Soler, C., Fabre, M., Vong, R., Dehan, C., Cazajous, G., Massoure, P.L., Kraemer, P., Jenkins, A., Garnotel, E., Pourcel, C., Vergnaud, G., 2012. Significance of the identification in the horn of africa of an exceptionally deep branching *Mycobacterium tuberculosis* Clade. PLoS One 7, e52841. https://doi.org/10.1371/journal.pone.0052841.

Brudey, K., Driscoll, J.R., Rigouts, L., Prodinger, W.M., Gori, A., Al-Hajoj, S.A., Allix, C., Aristimuño, L., Arora, J., Baumanis, V., Binder, L., Cafrune, P., Cataldi, A., Cheong, S., Diel, R., Ellermeier, C., Evans, J.T., Fauville-Dufaux, M., Ferdinand, S., Garcia de Viedma, D., Garzelli, C., Gazzola, L., Gomes, H.M., Guttierez, M.C., Hawkey, P.M., van Helden, P.D., Kadival, G.V., Kreiswirth, B.N., Kremer, K., Kubin, M., Kulkarni, S.P., Liens, B., Lillebaek, T., Ho, M.L., Martin, C., Martin, C., Mokrousov, I., Narvskaïa, O., Ngeow, Y.F., Naumann, L., Niemann, S., Parwati, I., Rahim, Z., Rasolofo-Razanamparany, V., Rasolonavalona, T., Rossetti, M.L., Rüsch-Gerdes, S., Sajduda, A., Samper, S., Shemyakin, I.G., Singh, U.B., Somoskovi, A., Skuce, R.A., van Soolingen, D., Streicher, E.M., Suffys, P.N., Tortoli, E., Tracevska, T., Vincent, V., Victor, T.C., Warren, R.M., Yap, S.F., Zaman, K., Portaels, F., Rastogi, N., Sola, C., 2006. *Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification population genetics

and epidemiology. BMC Microbiol. 6, 23. https://doi.org/10.1186/1471-2180-6-23.
Chernyaeva, E.N., Shulgina, M.V., Rotkevich, M.S., Dobrynin, P.V., Simonov, S.A., Shitikov, E.A., Ischenko, D.S., Karpova, I.Y., Kostryukova, E.S., Ilina, E.N., Govorun, V.M., Zhuravlev, V.Y., Manicheva, O.A., Yablonsky, P.K., Isaeva, Y.D., Nosova, E.Y., Mokrousov, I.V., Vyazovaya, A.A., Narvskaya, O.V., Lapidus, A.L., O'Brien, S.J., 2014. Genome-wide *Mycobacterium tuberculosis* variation (GMTV) database: a new tool for integrating sequence variations and epidemiology. BMC Genomics 15, 308. https://doi.org/10.1186/1471-2164-15-308.
Coll, F., McNerney, R., Guerra-Assunção, J.A., Glynn, J.R., Perdigão, J., Viveiros, M., Portugal, I., Pain, A., Martin, N., Clark, T.G., 2014a. A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. Nat. Commun. 5 (4812). https://doi.org/10.1038/ncomms5812.
Coll, F., Preston, M., Guerra-Assunção, J.A., Hill-Cawthorn, G., Harris, D., Perdigão, J., Viveiros, M., Portugal, I., Drobniewski, F., Gagneux, S., Glynn, J.R., Pain, A., Parkhill, J., McNerney, R., Martin, N., Clark, T.G., 2014b. PolyTB: a genomic variation map for *Mycobacterium tuberculosis*. Tuberculosis (Edinb.) 94, 346–354. https://doi.org/10.1016/j.tube.2014.02.005.
Comas, I., Homolka, S., Niemann, S., Gagneux, S., 2009. Genotyping of genetically monomorphic bacteria: DNA sequencing in *Mycobacterium tuberculosis* highlights the limitations of current methodologies. PLoS One 4, e7815. https://doi.org/10.1371/journal.pone.0007815.
Comas, I., Coscolla, M., Luo, T., Borrell, S., Holt, K.E., Kato-Maeda, M., Parkhill, J., Malla, B., Berg, S., Thwaites, G., Yeboah-Manu, D., Bothamley, G., Mei, J., Wei, L., Bentley, S., Harris, S.R., Niemann, S., Diel, R., Aseffa, A., Gao, Q., Young, D., Gagneux, S., 2013. Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. Nat. Genet. 45, 1176–1182. https://doi.org/10.1038/ng.2744.
Couvin, D., Rastogi, N., 2015. Tuberculosis – a global emergency: tools and methods to monitor, understand, and control the epidemic with specific example of the Beijing lineage. Tuberculosis (Edinb.) 95, S177–S189. https://doi.org/10.1016/j.tube.2015.02.023.
Couvin, D., Zozio, T., Rastogi, N., 2017. SpolSimilaritySearch - a web tool to compare and search similarities between spoligotypes of *Mycobacterium tuberculosis* complex. Tuberculosis (Edinb.) 105, 49–52. https://doi.org/10.1016/j.tube.2017.04.007.
Demay, C., Liens, B., Burguière, T., Hill, V., Couvin, D., Millet, J., Mokrousov, I., Sola, C., Zozio, T., Rastogi, N., 2012. SITVITWEB – a publicly available international multimarker database for studying *Mycobacterium tuberculosis* genetic diversity and molecular epidemiology. Infect. Genet. Evol. 12, 755–766. https://doi.org/10.1016/j.meegid.2012.02.004.
Dheda, K., Shean, K., Zumla, A., Badri, M., Streicher, E.M., Page-Shipp, L., Willcox, P., John, M.A., Reubenson, G., Govindasamy, D., Wong, M., Padanilam, X., Dziwiecki, A., van Helden, P.D., Siwendu, S., Jarand, J., Menezes, C.N., Burns, A., Victor, T., Warren, R., Grobusch, M.P., van der Walt, M., Kvasnovsky, C., 2010. Early treatment outcomes and HIV status of patients with extensively drug-resistant tuberculosis in South Africa: a retrospective cohort study. Lancet 375, 1798–1807. https://doi.org/10.1016/S0140-6736(10)60492-8.
Driscoll, J.R., Bifani, P.J., Mathema, B., McGarry, M.A., Zickas, G.M., Kreiswirth, B.N., Taber, H.W., 2002. Spoligologos: a bioinformatic approach to displaying and analyzing *Mycobacterium tuberculosis* data. Emerg. Infect. Dis. 8, 1306–1309. https://doi.org/10.3201/eid0811.020174.
Farmer, P.E., Nizeye, B., Stulac, S., Keshavjee, S., 2006. Review Structural violence and clinical medicine. PLoS Med. 3, e449. https://doi.org/10.1371/journal.pmed.0030449.
Ferdinand, S., Millet, J., Accipe, A., Cassadou, S., Chaud, P., Levy, M., Théodore, M., Rastogi, N., 2013. Use of genotyping based clustering to quantify recent tuberculosis transmission in Guadeloupe during a seven years period: analysis of risk factors and access to health care. BMC Infect. Dis. 13 (364). https://doi.org/10.1186/1471-2334-13-364.
Filliol, I., Driscoll, J.R., Van Soolingen, D., Kreiswirth, B.N., Kremer, K., Valétudie, G., Anh, D.D., Barlow, R., Banerjee, D., Bifani, P.J., Brudey, K., Cataldi, A., Cooksey, R.C., Cousins, D.V., Dale, J.W., Dellagostin, O.A., Drobniewski, F., Engelmann, G., Ferdinand, S., Gascoyne-Binzi, D., Gordon, M., Gutierrez, M.C., Haas, W.H., Heersma, H., Källenius, G., Kassa-Kelembho, E., Koivula, T., Ly, H.M., Makristathis, A., Mammina, C., Martin, G., Moström, P., Mokrousov, I., Narbonne, V., Narvskaya, O., Nastasi, A., Niobe-Eyangoh, S.N., Pape, J.W., Rasolofo-Razanamparany, V., Ridell, M., Rossetti, M.L., Stauffer, F., Suffys, P.N., Takiff, H., Texier-Maugein, J., Vincent, V., De Waard, J.H., Sola, C., Rastogi, N., 2002. Global distribution of *Mycobacterium tuberculosis* spoligotypes. Emerg. Infect. Dis. 8, 1347–1349. https://doi.org/10.3201/eid0811.020125.
Filliol, I., Driscoll, J.R., van Soolingen, D., Kreiswirth, B.N., Kremer, K., Valétudie, G., Dang, D.A., Barlow, R., Banerjee, D., Bifani, P.J., Brudey, K., Cataldi, A., Cooksey, R.C., Cousins, D.V., Dale, J.W., Dellagostin, O.A., Drobniewski, F., Engelmann, G., Ferdinand, S., Gascoyne-Binzi, D., Gordon, M., Gutierrez, M.C., Haas, W.H., Heersma, H., Kassa-Kelembho, E., Ho, M.L., Makristathis, A., Mammina, C., Martin, G., Moström, P., Mokrousov, I., Narbonne, V., Narvskaya, O., Nastasi, A., Niobe-Eyangoh, S.N., Pape, J.W., Rasolofo-Razanamparany, V., Ridell, M., Rossetti, M.L., Stauffer, F., Suffys, P.N., Takiff, H., Texier-Maugein, J., Vincent, V., de Waard, J.H., Sola, C., Rastogi, N., 2003. Snapshot of moving and expanding clones of *Mycobacterium tuberculosis* and their global distribution assessed by spoligotyping in an international study. J. Clin. Microbiol. 41, 1963–1970. https://doi.org/10.1128/JCM.41.5.1963-1970.2003.
Filliol, I., Motiwala, A.S., Cavatore, M., Qi, W., Hazbón, M.H., Bobadilla del Valle, M., Fyfe, J., García-García, L., Rastogi, N., Sola, C., Zozio, T., Guerrero, M.I., León, C.I., Crabtree, J., Angiuoli, S., Eisenach, K.D., Durmaz, R., Joloba, M.L., Rendón, A., Sifuentes-Osornio, J., Ponce de León, A., Cave, M.D., Fleischmann, R., Whittam, T.S., Alland, D., 2006. 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. 188, 759–772. https://doi.org/10.1128/JB.188.2.759-772.2006.
Firdessa, R., Berg, S., Hailu, E., Schelling, E., Gumi, B., Erenso, G., Gadisa, E., Kiros, T., Habtamu, M., Hussein, J., Zinsstag, J., Robertson, B.D., Ameni, G., Lohan, A.J., Loftus, B., Comas, I., Gagneux, S., Tschopp, R., Yamuah, L., Hewinson, G., Gordon, S.V., Young, D.B., Aseffa, A., 2013. Mycobacterial lineages causing pulmonary and extrapulmonary tuberculosis, Ethiopia. Emerg. Infect. Dis. 19, 460–463. https://doi.org/10.3201/eid1903.120256.
Gagneux, S., DeRiemer, K., Van, T., Kato-Maeda, M., de Jong, B.C., Narayanan, S., Nicol, M., Niemann, S., Kremer, K., Gutierrez, M.C., Hilty, M., Hopewell, P.C., Small, P.M., 2006. Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. Proc. Natl. Acad. Sci. U. S. A. 103, 2869–2873. https://doi.org/10.1073/pnas.0511240103.
García de Viedma, D., Mokrousov, I., Rastogi, N., 2011. Innovations in the molecular epidemiology of tuberculosis. Enferm. Infecc. Microbiol. Clin. 29, 8–13. https://doi.org/10.1016/S0213-005X(11)70012-X.
Hill, V., Zozio, T., Sadikalay, S., Viegas, S., Streit, E., Kallenius, G., Rastogi, N., 2012. MLVA based classification of *Mycobacterium tuberculosis* complex lineages for a robust phylogeographic snapshot of its worldwide molecular diversity. PLoS One 7, e41991. https://doi.org/10.1371/journal.pone.0041991.
Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., Hill, D.P., Kania, R., Schaeffer, M., St Pierre, S., Twigger, S., White, O., Rhee, S.Y., 2008. Big data: the future of biocuration. Nature 455, 47–50. https://doi.org/10.1038/455047a.
Ismail, F., Couvin, D., Farakhin, I., Abdul Rahman, Z., Rastogi, N., Suraiya, S., 2014. Study of *Mycobacterium tuberculosis* complex genotypic diversity in Malaysia reveals a predominance of ancestral East-African-Indian lineage with a Malaysia-specific signature. PLoS One 9, e114832. https://doi.org/10.1371/journal.pone.0114832.
Iwamoto, T., Grandjean, L., Arikawa, K., Nakanishi, N., Caviedes, L., Coronel, J., Sheen, P., Wada, T., Taype, C.A., Shaw, M.A., Moore, D.A., Gilman, R.H., 2012. Genetic diversity and transmission characteristics of Beijing family strains of *Mycobacterium tuberculosis* in Peru. PLoS One 7, e49651. https://doi.org/10.1371/journal.pone.0049651.
Jagielski, T., Minias, A., Van Ingen, J., Rastogi, N., Brzostek, A., Żaczek, A., Dziadek, J., 2016. Methodological and clinical aspects of the molecular epidemiology of *Mycobacterium tuberculosis* and other mycobacteria. Clin. Microbiol. Rev. 29, 239–290. https://doi.org/10.1128/CMR.00055-15.
Joshi, K.R., Dhiman, H., Scaria, V., 2014. tbvar: a comprehensive genome variation resource for *Mycobacterium tuberculosis*. Database bat083. https://doi.org/10.1093/database/bat083.
Kamerbeek, J., Schouls, L., Kolk, A., van Agterveld, M., van Soolingen, D., Kuijper, S., Bunschoten, A., Molhuizen, H., Shaw, R., Goyal, M., van Embden, J., 1997. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. J. Clin. Microbiol. 35, 907–914.
Keshavjee, S., Gelmanova, I.Y., Pasechnikov, A.D., Mishustin, S.P., Andreev, Y.G., Yedilbayev, A., Furin, J.J., Mukherjee, J.S., Rich, M.L., Nardell, E.A., Farmer, P.E., Kim, J.Y., Shin, S.S., 2008. Treating multidrug-resistant tuberculosis in Tomsk, Russia: developing programs that address the linkage between poverty and disease. Ann. N. Y. Acad. Sci. 1136, 1–11. https://doi.org/10.1196/annals.1425.009.
Kisa, O., Tarhan, G., Gunal, S., Albay, A., Durmaz, R., Saribas, Z., Zozio, T., Alp, A., Ceyhan, I., Tombak, A., Rastogi, N., 2012. Distribution of spoligotyping defined genotypic lineages among drug-resistant *Mycobacterium tuberculosis* complex clinical isolates in Ankara, Turkey. PLoS One 7, e30331. https://doi.org/10.1371/journal.pone.0030331.
Kohl, T.A., Diel, R., Harmsen, D., Rothgänger, J., Walter, K.M., Merker, M., Weniger, T., Niemann, S., 2014. Whole-genome-based *Mycobacterium tuberculosis* surveillance: a standardized, portable, and expandable approach. J. Clin. Microbiol. 52, 2479–2486. https://doi.org/10.1128/JCM.00567-14.
Koro Koro, F., Kamdem Simo, Y., Piam, F.F., Noeske, J., Gutierrez, C., Kuaban, C., Eyangoh, S.I., 2013. Population dynamics of tuberculous Bacilli in Cameroon as assessed by spoligotyping. J. Clin. Microbiol. 51, 299–302. https://doi.org/10.1128/JCM.01196-12.
Kremer, K., van Soolingen, D., Frothingham, R., Haas, W.H., Hermans, P.W., Martín, C., Palittapongarnpim, P., Plikaytis, B.B., Riley, L.W., Yakrus, M.A., Musser, J.M., van Embden, J.D., 1999. Comparison of methods based on different molecular epidemiological markers for typing of *Mycobacterium tuberculosis* complex strains: interlaboratory study of discriminatory power and reproducibility. J. Clin. Microbiol. 37, 2607–2618.
Lanzas, F., Karakousis, P.C., Sacchettini, J.C., Ioerger, T.R., 2013. Multidrug-resistant tuberculosis in Panama is driven by clonal expansion of a multidrug-resistant *Mycobacterium tuberculosis* strain related to the KZN extensively drug-resistant M. tuberculosis strain from South Africa. J. Clin. Microbiol. 51, 3277–3285. https://doi.org/10.1128/JCM.01122-13.
Millet, J., Streit, E., Berchel, M., Bomer, A.G., Schuster, F., Paasch, D., Vanhomwegen, J., Cadelis, G., Rastogi, N., 2014. A systematic follow-up of *Mycobacterium tuberculosis* drug-resistance and associated genotypic lineages in the French Departments of the Americas over a seventeen-year period. Biomed. Res. Int. 2014 (689852). https://doi.org/10.1155/2014/689852.
Mokrousov, I., 2012. The quiet and controversial: Ural family of *Mycobacterium tuberculosis*. Infect. Genet. Evol. 12, 619–629. https://doi.org/10.1016/j.meegid.2011.09.026.
Mokrousov, I., 2018. On sunspots, click science and molecular iconography. Tuberculosis (Edinb.) 110, 91–95. https://doi.org/10.1016/j.tube.2018.04.004.
Mokrousov, I., Jiao, W.W., Wan, K., Shen, A., 2014. Stranger in a strange land: Ibero-American strain of *Mycobacterium tuberculosis* in Tibet, China. Infect. Genet. Evol. 26, 323–326. https://doi.org/10.1016/j.meegid.2014.06.011.

Oelemann, M.C., Diel, R., Vatin, V., Haas, W., Rüsch-Gerdes, S., Locht, C., Niemann, S., Supply, P., 2007. Assessment of an optimized mycobacterial interspersed repetitive-unit-variable-number tandem-repeat typing system combined with spoligotyping for population-based molecular epidemiology studies of tuberculosis. J. Clin. Microbiol. 45, 691–697. https://doi.org/10.1128/JCM.01393-06.

Perdigão, J., Silva, C., Diniz, J., Pereira, C., Machado, D., Ramos, J., Silva, H., Abilleira, F., Brum, C., Reis, A.J., Macedo, M., Scaini, J.L., Silva, A.B., Esteves, L., Macedo, R., Maltez, F., Clemente, S., Coelho, E., Viegas, S., Rabna, P., Rodrigues, A., Taveira, N., Jordao, L., Kritski, A., Lapa, E., Silva, J.R., Mokrousov, I., Couvin, D., Rastogi, N., Couto, I., Pain, A., McNerney, R., Clark, T.G., von Groll, A., Dalla-Costa, E.R., Rossetti, M.L., Silva, P.E.A., Viveiros, M., Portugal, I., 2018. Clonal expansion across the seas as seen through CPLP-TB database: a joint effort in cataloguing *Mycobacterium tuberculosis* genetic diversity in Portuguese-speaking countries. Infect. Genet. Evol. https://doi.org/10.1016/j.meegid.2018.03.011. In press.

Rastogi, N., Sola, C., 2007. Chapter 2 - Molecular evolution of the *Mycobacterium tuberculosis* complex. In Tuberculosis 2007: from basic science to patient care. In: Palomino, J.C., Leao, S., Ritacco, V. (Eds.), Amedeo Online Textbooks, pp. 53–91. http://pdf.flyingpublisher.com/tuberculosis2007.pdf.

Reyes, J.F., Francis, A.R., Tanaka, M.M., 2008. Models of deletion for visualizing bacterial variation: an application to tuberculosis spoligotypes. BMC Bioinformatics 9, 496. https://doi.org/10.1186/1471-2105-9-496.

Rodriguez-Campos, S., González, S., de Juan, L., Romero, B., Bezos, J., Casal, C., Álvarez, J., Fernández-de-Mera, I.G., Castellanos, E., Mateos, A., Sáez-Llorente, J.L., Domínguez, L., Aranaz, A., Spanish Network on Surveillance Monitoring of Animal Tuberculosis, 2012. A database for animal tuberculosis (mycoDB.Es) within the context of the Spanish national program for eradication of bovine tuberculosis. Infect. Genet. Evol. 12, 877–882. https://doi.org/10.1016/j.meegid.2011.10.008.

Rothschild, B.M., Martin, L.D., Lev, G., Bercovier, H., Bar-Gal, G.K., Greenblatt, C., Donoghue, H., Spigelman, M., Brittain, D., 2001. *Mycobacterium tuberculosis* complex DNA from an extinct bison dated 17,000 years before the present. Clin. Infect. Dis. 33, 305–311. https://doi.org/10.1086/321886.

Shabbeer, A., Cowan, L.S., Ozcaglar, C., Rastogi, N., Vandenberg, S.L., Yener, B., Bennett, K.P., 2012. TB-Lineage: an online tool for classification and analysis of strains of *Mycobacterium tuberculosis* complex. Infect. Genet. Evol. 12, 789–797. https://doi.org/10.1016/j.meegid.2012.02.010.

Smith, N.H., Upton, P., 2012. Naming spoligotype patterns for the RD9-deleted lineage of the *Mycobacterium tuberculosis* complex; www.Mbovis.org. Infect. Genet. Evol. 12, 873–876. https://doi.org/10.1016/j.meegid.2011.08.002.

Soares, P., Alves, R.J., Abecasis, A.B., Penha-Gonçalves, C., Gomes, M.G., Pereira-Leal, J.B., 2013. inTB - a data integration platform for molecular and clinical epidemiological analysis of tuberculosis. BMC Bioinformatics 14, 264. https://doi.org/10.1186/1471-2105-14-264.

Sola, C., Filliol, I., Gutierrez, M.C., Mokrousov, I., Vincent, V., Rastogi, N., 2001. Spoligotype database of *Mycobacterium tuberculosis*: biogeographic distribution of shared types and epidemiologic and phylogenetic perspectives. Emerg. Infect. Dis. 7, 390–396. https://doi.org/10.3201/eid0703.010304.

Stucki, D., Gagneux, S., 2013. Single nucleotide polymorphisms in *Mycobacterium tuberculosis* and the need for a curated database. Tuberculosis (Edinb.) 93, 30–39. https://doi.org/10.1016/j.tube.2012.11.002.

Supply, P., Allix, C., Lesjean, S., Cardoso-Oelemann, M., Rusch-Gerdes, S., Willery, E., Savine, E., de Haas, P., van Deutekom, H., Roring, S., Bifani, P., Kurepina, N., Kreiswirth, B., Sola, C., Rastogi, N., Vatin, V., Gutierrez, M.C., Fauville, M., Niemann, S., Skuce, R., Kremer, K., Locht, C., van Soolingen, D., 2006. Proposal for Standardization of Optimized Mycobacterial Interspersed Repetitive Unit-Variable-Number Tandem repeat Typing of *Mycobacterium tuberculosis*. J. Clin. Microbiol. 44, 4498–4510. https://doi.org/10.1128/JCM.01392-06.

Tang, C., Reyes, J.F., Luciani, F., Francis, A.R., Tanaka, M.M., 2008. SpolTools: online utilities for analyzing spoligotypes of the *Mycobacterium tuberculosis* complex. Bioinformatics 24, 2414–2415. https://doi.org/10.1093/bioinformatics/btn434.

Tulu, B., Ameni, G., 2018. Spoligotyping based genetic diversity of *Mycobacterium tuberculosis* in Ethiopia: a systematic review. BMC Infect. Dis. 18 (140). https://doi.org/10.1186/s12879-018-3046-4.

van Embden, J.D., Cave, M.D., Crawford, J.T., Dale, J.W., Eisenach, K.D., Gicquel, B., Hermans, P., Martin, C., McAdam, R., Shinnick, T.M., Small, P.M., 1993. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. J. Clin. Microbiol. 31, 406–409.

van Soolingen, D., 2014. Whole-genome sequencing of *Mycobacterium tuberculosis* as an epidemiological marker. Lancet Respir. Med. 2, 251–252. https://doi.org/10.1016/S2213-2600(14)70049-9.

Velayati, A.A., Masjedi, M.R., Farnia, P., Tabarsi, P., Ghanavi, J., ZiaZarifi, A.H., Hoffner, S.E., 2009. Emergence of new forms of totally drug-resistant tuberculosis bacilli: super extensively drug-resistant tuberculosis or totally drug-resistant strains in Iran. Chest 136, 420–425. https://doi.org/10.1378/chest.08-2427.

Weniger, T., Krawczyk, J., Supply, P., Niemann, S., Harmsen, D., 2010. MIRU-VNTRplus: a web tool for polyphasic genotyping of *Mycobacterium tuberculosis* complex bacteria. Nucleic Acids Res. 38, W326–W331. https://doi.org/10.1093/nar/gkq351.

World Health Organization (WHO), 2017. Global tuberculosis report. pp. 2017. http://www.who.int/tb/publications/global_report/gtbr2017_main_text.pdf.

Yuan, X., Zhang, T., Kawakami, K., Zhu, J., Zheng, W., Li, H., Deng, G., Tu, S., Liu, W., 2013. Genotyping and clinical characteristics of multidrug and extensively drug-resistant tuberculosis in a tertiary care tuberculosis hospital in China. BMC Infect. Dis. 13 (315). https://doi.org/10.1186/1471-2334-13-315.

Zignol, M., Hosseini, M.S., Wright, A., Weezenbeek, C.L., Nunn, P., Watt, C.J., Williams, B.G., Dye, C., 2006. Global incidence of multidrug-resistant tuberculosis. J. Infect. Dis. 194, 479–485. https://doi.org/10.1086/505877.

Zink, A.R., Sola, C., Reischl, U., Grabner, W., Rastogi, N., Wolf, H., Nerlich, A.G., 2003. Characterization of *Mycobacterium tuberculosis* complex DNAs from Egyptian mummies by spoligotyping. J. Clin. Microbiol. 41, 359–367.