EuroReference
journal of Reference
anses
French agency for food, environmental and occupational health & safety

Journal No. 12
Summer 2014

Research

The establishment of databases on circulating genotypes of *Mycobacterium tuberculosis* complex and web tools for an effective response to better monitor, understand and control the tuberculosis epidemic worldwide

David Couvin et Nalin Rastogi (nrastogi@pasteur-guadeloupe.fr)
WHO Supranational TB Reference Laboratory, Institut Pasteur de la Guadeloupe, 97183 Abymes, Guadeloupe, France.

Key words: *Mycobacterium tuberculosis*, tuberculosis, genotyping, databases, spoligotyping, MIRU-VNTR, epidemiology, phylogeny, drug-resistance.

Abstract

In this paper we will briefly review various *Mycobacterium tuberculosis* complex (MTBC) genotyping databases developed over last fifteen years at the Institut Pasteur de la Guadeloupe, which represent a great concerted initiative and effort to control the tuberculosis epidemic. Starting from the initial excel-based version in 1999 (SpolDB1; n=610 clinical isolates) to the fourth MySQL-based version in 2006 (SpolDB4; n=39,295 clinical isolates), these databases permitted to have a first phylogeographical snapshot of circulating MTBC genotypic lineages based on spacer oligonucleotide typing (spoligotyping) which allows to study the polymorphism of the Direct Repeat (DR) locus. The two most recent MySQL-based multimarker versions concern the 5th version SITVITWEB released in 2012 (n=62,582 clinical isolates) with both spoligotyping and 12-loci Mycobacterial Interspersed Repetitive Units - Variable-Number of Tandem Repeats (MIRU-VNTRs); and the 6th version named SITVIT2 that will be released in 2014 (n= 111,635 clinical isolates) with spoligotyping and 12-, 15- or 24-loci MIRU-VNTR data. In these recent versions, a web-based interface allows the user to search for strains through the database by criteria, such as the year, the isolation country, the country of origin, the investigator's name. It further facilitates to perform combined searches in SITVIT2, making it possible to get the genotyping data on selected strains in conjunction with their geographical distribution, as well as available data on drug-resistance, demographic and epidemiologic characteristics. Our research initiative is thus focused to further 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 and regional 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.

Introduction

Almost twenty years after the World Health Organization (WHO) declaration of tuberculosis (TB) as a global public health emergency, and despite the major progress made towards 2015 global targets set within the context of the millennium development goals, TB is still the second deadliest disease caused by an infectious agent in the world after HIV/AIDS (WHO Report, 2013). In 2012, it led to an estimated 8.6 million new cases and 1.3 million deaths (including 320,000 deaths among TB/HIV co infected patients). A careful scrutiny of the WHO Report shows that TB remains an enormous health and economic problem not only in developing countries but also in developed nations due to the TB-HIV co-infection and emergence of multidrug-resistant (MDR), and more recently of extremely drug-resistant (XDR) isolates, which further complicates management of the disease and considerably increases the mortality due to TB among immune-compromised patients.

Furthermore, the increasing rate of travel/migration of population for leisure or work in the last decades has led to a new challenge in countries where TB was declining due to changes in the socio-epidemiological scenarios generated by massive immigration from countries where TB is highly endemic (García de Viedma *et al.,* 2011). Some key questions include comparison of the role of recent transmission with that of reactivation/importation in TB among foreign-born cases, the impact of potential importation of previously unidentified *M. tuberculosis* strains, and cross-transmission between cases from different nationalities. It is therefore important to understand how tubercle bacilli are transmitted, which clones are involved in drug-resistant cases and/or outbreaks, identify new clones that may be emerging in a setting vs. those that may be under extinction, identify subpopulations and risk factors with highest threat of catching the infection, and be able to interpret these results within the evolutionary framework of *M. tuberculosis* complex (MTBC). It is clear today that these questions would prove difficult to answer without the support of molecular epidemiology.

Until recently, all human-adapted strains of MTBC were traditionally considered to be essentially identical, hence the question of individual genetic variation within MTBC gained little attention which led to most previous research being focused on the individual organism. The advent of molecular methods and their widespread use in population-based studies introduced both new conceptual and technological developments. Although, MTBC constitutes a remarkably homogeneous group genetically with proven evidence of clonal evolution, recent studies have shown that the genetic diversity among individual clones is much higher than previously assumed, with potential impact on pathobiological properties. A leading study on a global collection of MTBC strains using seven megabase pairs of DNA sequence data showed significant genetic diversity

Research

(Hershberg *et al.,* 2008). The authors suggested that much of this genetic diversity was driven by genetic drift potentially linked to human demographic and migratory events with functional consequences such as emergence and spread of drug-resistant tuberculosis.
Although TB is present worldwide, some countries/regions contain higher incidences than others: e.g. Sub-Saharan Africa, Southern and Southeastern Asia, Latin America, but also Eastern Europe and Russia, as well as the Caribbean (particularly Haiti and the Dominican Republic). Indeed, TB still constitutes a major health problem in many of the Caribbean and Latin-American countries; e.g., the incidence of TB in Haiti which was 330/100,000 in the 90s is still as high as 213/100,000 in 2012 (WHO Report, 2013). It is obvious that early diagnosis and bacteriological confirmation of TB, and drug-resistance determination is a key process to control the TB epidemic, and reference laboratories represent essential structures in the global diagnosis and control of the disease worldwide. In such a context, Institut Pasteur took the initiative to start the first TB reference laboratory for the Caribbean region in 1993. Situated at Institut Pasteur de la Guadeloupe (IPG), it was initially targeted to work for the Caribbean; nonetheless being well aware of TB's global context, we started right from the beginning with an approach based on concomitant use of bacteriological and molecular methods. Over the last 2 decades, we developed a comprehensive approach which deals with routine and molecular diagnostics of MTBC and other mycobacterial species, drug-resistance surveillance, and development of rapid diagnostic methods and diverse molecular techniques that are useful for epidemiological and population-based studies at local, regional, and global level. For global TB surveillance, we developed a series of TB genotyping databases since 1999 which accumulated our own data as well as collected from various participating laboratories worldwide. This report will briefly summarize the steps undertaken to develop these databases and the web-based tools developed to offer the possibility of creating worldwide geographical distribution maps displaying the frequencies of TB genotypes worldwide at various geographical scales. We will also briefly refer to some of the published or ongoing studies making use of this information and future prospects.

TB: origin, spread and co-adaptation with its hosts

With evidence of the isolation and characterization of ancient *M. tuberculosis* DNA from an extinct bison dated 17,000 years B.C., suggesting the presence of TB in America in the late Pleistocene (Rothschild *et al.,* 2001), TB was already known as a very ancient disease. When looking at human remains, ancient DNA helped to trace the presence of TB in Egyptian mummies, with characterization of *M. tuberculosis* and *M. africanum* (Zink *et al.,* 2003). Subsequently, in analogy to other crowd diseases, the origin of human tuberculosis was thought to be associated with the Neolithic Demographic Transition (NDT) starting around 11,000 years ago, as the development of animal domestication increased the likelihood of zoonotic transfer of novel pathogens to humans while agricultural innovations supported increased population densities that helped sustain the infectious cycle (Wolfe *et al.,* 2007). However, it was not clear whether (a) human TB descended from a ruminant mycobacterium that recently infected humans from domestic animals, or from an ancient human mycobacterium that has come to infect domestic and wild ruminants; and (b) whether tuberculosis originated independently in both hemispheres or was brought to the Americas by Europeans. Nonetheless, TB also displayed a pattern of chronic progression, latency and reactivation which is characteristic of a pre-NDT disease (Barry *et al.,* 2009).
The answer to these questions came thanks to new research which showed that TB is probably as old as humanity itself (Comas *et al.,* 2013). By studying the diverse genetic variations in MTBC, the researchers were able to show that TB must have spread around the world with the first modern humans to emerge from Africa. This study analyzed the whole genomes of a collection of 259 contemporary strains of MTBC from around the world, and compared MTBC phylogenic diversity to human diversity inferred from mitochondrial genome data. The results indicated that MTBC emerged about 70,000 years ago, accompanied migrations of anatomically modern humans out of Africa, and expanded as a consequence of increases in human population density during the Neolithic period (Comas *et al.,* 2013). This early origin of TB and the fact that genome-based phylogeny of MTBC mirrored that of human mitochondrial genomes, further showed that TB lung infection did not spread to humans from their domesticated animals since farming came much later. This landmark study also showed striking similarities in the evolutionary path of humans and MTBC, and suggested that MTBC evolution not only paralleled that of humans but also that MTBC diversity directly benefited from human demographic explosions. The fact that *M. tuberculosis* "stricto-sensu" is an obligate human pathogen with no known animal or environmental reservoir, changes in human demography and population densities over time are most likely to affect its evolution like in case of a crowd disease model. At the same time, its latency and chronicity possibly allows it to adapt to lower host densities, survive, and strike back when favorable conditions allow massive host infections.

Typing methods for TB molecular epidemiology

MTBC is a very diverse group of organisms ecologically, and includes *M. tuberculosis, M. africanum*, and *M. canettii* (exclusively human pathogens), *M. microti* (a rodent pathogen), and *M. bovis* (bovine pathogen but with a wide host range), as well as *M. pinnipedii* (seals), *M. caprae* (goats), *M. mungi* (banded Mongooses), and the oryx (renamed *M. orygis*), dassie, and chimpanzee bacilli, causative agents of TB in the animal species after which they are named. However, MTBC constitutes a highly homogeneous group genetically and various MTBC members share on average more than 99.7 % of nucleotide identity (Kato-Maeda *et al.,* 2001). Despite this remarkable genetic homogeneity, the last 2 decades have witnessed the advent of new molecular methods widely used today in population-based genotyping studies, permitting to precisely characterize TB isolates, and infer different phylogenetic lineages associated. Detailed reviews are available on MTBC molecular evolution (Rastogi and Sola, 2007), current molecular typing methods (Jagielski *et al.,* 2014), and strategies and innovations in the broad field of TB molecular epidemiology (2) [García de Viedma *et al.,* 2011), and interested readers are referred to these for detailed information.
Although the IS6110-RFLP method was long considered the gold standard technique for *M. tuberculosis* typing due to its reproducibility and discriminatory power in the molecular epidemiological investigations of TB (van Embden *et al.,* 1993),

Research

this labor-intensive methodology required large quantities of DNA, and was characterized by a lack of discriminatory power for typing of the isolates with low copy numbers of IS6110, e.g. in South India (Radhakrishnan *et al.,* 2001). Furthermore, the fast molecular clock of this marker for evolutionary studies (de Boer *et al.,* 1999), the complexity of forces driving its transposition and risk of genetic convergence (Fang *et al.,* 2001), and difficulty to build large RFLP databases and need for sophisticated software for data analysis (Heersma *et al.,* 1998; Salamon *et al.,* 1998), rendered its use in evolutionary genetics of limited interest. For the reasons mentioned above, alternative PCR-based typing strategies such as Spoligotyping (Kamerbeek *et al.,* 1997) and mycobacterial interspersed repetitive units-variable number of tandem repeats or MIRU-VNTRs (Supply *et al.,* 2001; 2006) largely replaced IS6110-RFLP in the last decade; thus creating the basis for large-scale, high-throughput *M. tuberculosis* genotyping.

Based on the polymorphism of the direct repeat (DR) locus, Spoligotyping for "**SP**acer **OLIGO** nucleotide **TYPING**" is currently one of the most frequently used PCR-based approaches for studying the molecular epidemiology and phylogeography of MTBC. Initially identified in the vaccine strain M. bovis BCG (Hermans *et al.,* 1991), the DR locus belongs to the CRISPR ("clustered regularly interspaced short palindromic repeat") family of repetitive DNAs, and contains multiple identical 36-bp perfect repeats (3 helix turns) interspersed by unique 34-41 bp spacers. Together with their non-repetitive spacer sequences, the DR units constitute multiple direct variant repeats (DVRs) which show extensive polymorphism among *M. tuberculosis* clinical isolates. The 43 character binary pattern generated by this technique was previously shown to vehicle significant phylogenetic information (Sola *et al.,* 2001). Spoligotype patterns are now commonly designated with their octal description, an internationally-agreed reporting format (Dale *et al.,* 2001). Because of its simplicity, binary result format and high reproducibility, spoligotyping is widely used for investigations on MTBC molecular epidemiology as a macroarray-based method to study presence/absence of 43 selected spacers (out of 104 spacers present). Indeed, a PubMed search for "spoligotyping OR spoligotype" gave 924 published articles (interrogation made on 18 March 2014).

One of the limitations of Spoligotyping being its tendency to overestimate MTBC clustering, it was soon proposed to complement this method with a minisatellite (MIRU-VNTR) based typing in a «two PCR-based» strategy, in conjunction to conventional epidemiological investigations (Sola *et al.,* 2003). The MIRU-VNTR minisatellites constitute a multi-locus marker set since they represent independent markers of a same type, and have been used as classical 12-locus, discriminatory 15-locus, or full 24-locus formats (Supply *et al.,* 2001; 2006). However, even the 24-locus MIRUs lacked a satisfactory resolution power for accurately discriminating closely related Beijing genotype strains, a fact that led to a recent proposal to use an additional 4-locus set of consensus hypervariable MIRU-VNTRs for subtyping Beijing clonal complexes and clusters (Allix-Béguec *et al.,* 2014).

Molecular typing methods for TB phylogeny

Apart from their use in epidemiology (García de Viedma *et al.,* 2011), molecular typing methods are also useful for evolutionary studies (Rastogi and Sola, 2007). These essentially include 2 sets of markers: (a) those summarized above which are extensively used for epidemiological studies but also provide with concomitant phylogenetical information – IS6110-RFLP, spoligotyping, and MIRU-VNTRs (see below); and (b) a set of markers including Large Sequence Polymorphisms (LSP) / Regions of Difference (RD), and Single Nucleotide Polymorphisms (SNPs), that are specifically useful for phylogenetical and evolutionary studies.

One of the earliest studies used substractive genomic hybridization to identify three distinct genomic regions between virulent *M. bovis, M. tuberculosis*, and the avirulent *M. bovis* BCG strain, designated respectively as RD1, RD2, and RD3 (Mahairas *et al.,* 1996). In another study, a distinction between three genetic groups of *M. tuberculosis* was achieved based on two polymorphisms occurring at high frequency in the genes encoding catalase-peroxidase and the A subunit of gyrase, which led to a classification in three principal genetic groups (PGG); group 1 bacteria being ancestral to groups 2 and 3 (Sreevatsan *et al.,* 1997). Almost immediately thereafter, restriction-digested bacterial artificial chromosome (BAC) arrays of H37Rv strain were used to reveal the presence of 10 regions of difference between *M. tuberculosis* and *M. bovis* (RD1 to 10); 7 of which (RD4-RD10) were deleted in *M. bovis* (Gordon *et al.,* 1999). In a major contribution, Brosch *et al.* (2002) analyzed the distribution of 20 variable regions resulting from insertion-deletion events in the genome of the tubercle bacilli in a collection of strains belonging to all MTBC subspecies, and showed that based on the presence or absence of a M. tuberculosis specific deletion 1 (TbD1, a 2 kb sequence), M. tuberculosis could be divided into "ancient" TbD1 positive and "modern" TbD1 negative strains (Brosch *et al.,* 2002). In this new evolutionary scenario of the *M. tuberculosis* complex, the RD9 deletion identifies an evolutionary lineage represented by *M. africanum, M. microti* and *M. bovis* that diverged from the progenitor of the present *M. tuberculosis* strains before TbD1 occurred, a finding which contradicts previous assumptions that *M. tuberculosis* evolved from a precursor of *M. bovis* (Brosch *et al.,* 2002). Since *M. canettii* and other ancestral *M. tuberculosis* complex strains lacked none of these regions, they are supposed to be direct descendants of the tubercle bacilli that existed before the "*M. africanum - M. bovis*" lineage separated from the *M. tuberculosis* lineage.

Using a global MTBC collection and 212 SNPs, Filliol *et al.* (2006) identified six deeply branching, phylogenetically distinct SNP cluster groups (SCGs) and five subgroups. The SCGs were strongly associated with the geographical origin of the *M. tuberculosis* samples and the birthplace of the human hosts. The authors proposed an algorithm able to identify two minimal sets of either 45 or 6 SNPs out of 212 SNPs tested, that could be used for screening of global MTBC collections for studies on evolution, strain differentiation, and biological differences among strains. In another study, Gutacker *et al.,* (2006) studied MTBC genetic relationships by analyzing 36 sSNPs among a big collection of strains from patients enrolled in 4 population-based studies in the United States and Europe, and assigned the strain collection to 1 of 9 major genetic clusters. A similar classification was revealed by analysis of other extended SNPs. Since the classification patterns of the SNP-based phylogenetic lineages were non-randomly associated with IS6110 profiles, spoligotypes, and MIRU-VNTRs, the authors argued for a strongly clonal MTBC population structure.

In parallel, by using DNA microarrays to comprehensively identify large-sequence polymorphisms, a stable association between

Research

MTBC strains and their human host populations was observed (Hirsh *et al.,* 2004); phylogenetic analysis not only indicated that horizontal gene transfers were rare among MTBC, but also that associations between host and pathogen populations were stable even in a cosmopolitan urban setting (like San Francisco), and were largely dictated by the composition of the local immigrant population. The authors concluded that *M. tuberculosis* is organized into several large, genetically differentiated populations, which in turn are directly and stably associated with host populations delineated according to their place of origin. A subsequent report by the same group confirmed this variable host-pathogen compatibility, the global *M. tuberculosis* population structure being defined by six RD/LSP-defined phylogeographical lineages – each associated with specific, sympatric human populations, i.e., the Indo-Oceanic lineage, East-Asian lineage, East-African-Indian lineage, Euro-American lineage, and two West-African lineages (Gagneux *et al.,* 2006.

Table 1. Comparison of spoligotyping-based nomenclature of *M. tuberculosis* lineages vs. PGG groupings, SNPs and SNP-based Cluster groups (SCG), and LSP-based lineages.

Spoligotyping-based (Filliol 2003)	PGG	SCG (Filliol 2006)	SNP-based (Gutacker 2006)	LSP (Gagneux 2006)
East-African-Indian (EAI)	PGG1	SCG 1	sSNP-I	Indo-Oceanic
Beijing	PGG1	SCG 2	sSNP-II	East-Asian
Central-Asian (CAS)	PGG1	SCG 3a	sSNP-IIA	East-African-Indian
Haarlem	PGG2	SCG 3b	sSNP-III	Euro-American
X1	PGG2	SCG 3c	sSNP-IV	Euro-American
X1,X2,X3	PGG2	SCG 4	sSNP-V	Euro-American
LAM	PGG2	SCG 5	sSNP-VI	Euro-American
T (Miscellaneous)	PGG2-3	SCG 6	sSNP-VII sSNP-VIII	Euro-American
Bovis	PGG1	SCG 7	(MTBC)	(MTBC)
M. africanum	PGG1	NA	NA	West-African 1
M. africanum	PGG1	NA	NA	West-African 2

Table 1 summarizes the correspondence among various lineage nomenclatures. It is important to underline that one must keep in mind the marker used when talking about a lineage, particularly for the naming of "East-African Indian" or EAI which denote 2 completely different groups of *M. tuberculosis* by spoligotyping vs. LSPs. Interestingly, a good congruence was observed between spoligotyping and SNPs (Filliol et al., 2006); the East African Indian and Beijing spoligotypes being concordant with SCG-1 and SCG-2, respectively; X and Central Asian spoligotypes were also associated with one SCG or subgroup combination. Other clades had less consistent associations with SCGs. Furthermore, the various spoligotyping-defined lineages fit well with the previous PGG groups, hence MTBC strains can be tentatively classified as ancestral TbD1+/PGG1 group (subset 1: M. africanum and East African Indian, EAI), modern TbD1– /PGG1 group (subset 2: Beijing and Central Asian or CAS), and evolutionary recent TbD1– /PGG2/3 group (subset 3: Haarlem, X, S, T, and Latin American and Mediterranean or LAM). Nonetheless, proper epidemiologic and phylogenetic inferences are not always an easy task due to a lack of understanding of the mechanisms behind the mutations leading to the polymorphism of these genomic targets. Recent studies have shown that phylogenetically unrelated MTBC strains could be sometimes found with the same spoligotype pattern as a result of independent mutational events (Fenner et al., 2011), an observation that corroborates the fact that spoligotyping is prone to homoplasy to a higher extent than the MIRU-VNTRs (Comas et al., 2009). Furthermore, spoligotyping has little discriminative power for families associated with the absence of large blocks of spacers, e.g., the Beijing lineage (Allix-Béguec et al., 2014). For all these reasons, we recommend to make a finer phylogenetic analysis of most significant circulating MTBC clones by multiple genetic markers and compare to the existing data worldwide – a complicated task by itself had it not been for availability of huge international databases that provide with such a framework today.

TB genotyping databases – what is available?

Our knowledge about TB is wider today than ever before, but what is our ability to compare the data generated with respect to all the data that has cumulated over years? Are we really able to instantaneously compare the genetic information on the circulating MTBC strains in conjunction with all demographical, clinical, bacteriological and epidemiological information available in various registers? The necessity of databases in such a context is obvious, and conception and design of databases in the control/surveillance of TB as well as other communicable diseases is certainly going to be an essential tool for achieving the Millennium Development Goals (MDGs) targeted by the WHO for 2015 (WHO, 2006). Indeed, databases allow the storing of huge amount of information in a structured way, facilitating data processing, interrogation, and streamline the decision process thanks to knowledge-based datamining. Nonetheless, they should be constantly updated and maintained like historical heritage and monuments in the present era of « Big Data », representing a more voluminous set of data exceeding the size of traditional databases, a fact which requires for revolutionary measures to be taken for data management, analysis and accessibility of biological data (Howe *et al.,* 2008). In the last couple of years, various databases and web tools have been developed in the TB field mostly devoted to study TB molecular epidemiology and evolution; some examples include:

- SpolDB4 and SITVITWEB are genotyping databases developed at IPG (Brudey *et al.,* 2006; Demay *et al.,* 2012). The later SITVITWEB version is a multimarker database with genotyping data on 62,582 clinical isolates corresponding to 153 countries of patient origin (105 countries of isolation). Method-wise it contains: (a) spoligotyping data, n=7105 patterns from 58180 clinical isolates, grouped in 2740 shared-types or SITs (n=53816 clinical isolates), and 4364 orphan patterns; (b) 12-locus MIRU-VNTRs, n=2379 patterns from 8161 clinical isolates, grouped in 847 shared-types or MITs (n=6626 clinical isolates), and 1533 orphan patterns; (c) 5-locus Exact Tandem Repeats (ETRs), n=458 patterns from 4626 clinical isolates, grouped in 245 shared-types or VITs (n=4413 clinical isolates), and 213 orphan patterns. The SITVITWEB database is freely available at: http://www.pasteur-guadeloupe.fr:8081/SITVIT_ONLINE
- SpolTools is a collection of online browser programs and visualization tool designed to manipulate and analyze MTBC spoligotyping data (Reyes *et al.,* 2008; Tang *et al.,* 2008). It

Research

also contains an online repository of spoligotyped isolates collected from published literature (currently 30 datasets containing 1179 spoligotype patterns corresponding to 6278 isolates). In particular, it allows to draw Spoligoforest trees that illustrate evolutionary relationships between spoligotypes in a given setting. SpolTools is available at: http://www.emi.unsw.edu.au/spolTools/

- MIRU-VNTRplus is a web based tool dedicated to analyze molecular typing data on TB, particularly the 12-, 15- and 24-loci MIRU-VNTRs formats (Allix-Béguec *et al.,* 2008; Weniger *et al.,* 2010). Tools for data exploration include search for similar strains, creation of phylogenetic and minimum spanning trees and mapping of geographic information. In addition, the database also provides detailed results (geographical origin, drug susceptibility profiles, genetic lineages and the spoligotyping pattern, SNP and LSP profiles, and the IS6110-RFLP fingerprints) on a collection of 186 well-characterized reference strains. MIRU-VNTRplus is available at: http://www.miru-vntrplus.org/
- TB Genotyping Information Management System (TB GIMS) is a secure web-based system designed to improve access and dissemination of genotyping information nationwide in the United States (44) [MMWR, 2010]. It stores and manages genotyping data on TB patients in the United States; allows authorized users to submit and track MTBC isolates to and from the contract genotyping labs; provides immediate notification of genotyping results and updates to TB labs and programs; links isolate data to patient-level surveillance data; provides reports on genotype clusters, including national genotype distribution; and provides national, state, and county maps of genotype clusters. This database is not publicly available.
- Mbovis.org is a spoligotype database with over 1400 patterns belonging to following RD9-deleted MTBC lineages: *M. africanum, M. bovis* (antelope), *M. microti, M. pinnipedii, M. caprae* and *M. bovis* (Smith and Upton, 2012). This database is available at: http://www.mbovis.org/
- MycoDB.es is a Spanish database of Animal tuberculosis (Rodriguez-Campos *et al.,* 2012), which was created as an epidemiological tool at national level (Spain). It contains 401 different spoligotype patterns containing 17,273 isolates belonging to *M. bovis, M. caprae* and *M. tuberculosis*, as well as a limited amount of MIRU-VNTR data. Unfortunately, this database is restricted to authorized access, limited to Spanish animal health agency – Centro de Vigilencia Sanitaria Veterinaria (VISAVET): http://www.vigilanciasanitaria.es/mycodb/
- TB-Lineage is an online tool for classification and analysis of MTBC genotypes into major lineages using spoligotypes and optionally MIRU locus 24 (Shabbeer *et al.,* 2012). It was developed and tested using genotyping data from the Centers for Disease Control and Prevention (CDC), Atlanta on 37066 clinical isolates corresponding to 3198 spoligotype patterns and 5430 MIRU-VNTR patterns. However, if MIRU locus 24 data is not available, the system utilizes predictions made by a Naïve Bayes classifier based on spoligotype data alone. The accuracy of automated classification using both spoligotypes and MIRU24 is >99%, and using spoligotypes alone is >95%. TB-Lineage is freely available at http://tbinsight.cs.rpi.edu/run_tb_lineage.html. This website also provides a tool to generate spoligoforests in order to visualize the genetic diversity and relatedness of genotypes and their associated lineages.
- tbvar is a searchable database using a systematic computational pipeline which allows to annotate potential functional and/or drug-resistance-associated variants from clinical re-sequencing data of MTBC (Joshi *et al.,* 2013). For this purpose, the authors re-analyzed re-sequencing datasets corresponding to more than 450 MTBC isolates available in public domain so as to create a comprehensive variome map comprising >29 000 single nucleotide variations. This database can be accessed by browsing location of variants (e.g., 1417019, 3037367, 4222628, etc.); genes (e.g., *katG, pncA, gyrA,* etc.); RvID (e.g., Rv1059, Rv1069c, Rv3693, etc.); or genome position range (10000-15000 ; 30000-35000 ; 80000-85000, etc.); and is available at: http://genome.igib.res.in/tbvar/
- InTB is a web-based interface/system for integrated warehousing and analysis of clinical, socio-demographic and molecular typing data on TB (Soares *et al.,* 2013). It allows to insert and download standard genotyping data in conjunction with an extensive array of clinical and socio-demographic variables that are used to characterize the disease. It also allows to classify new isolates into a well-characterized set of isolates based on internal references, multiple types of data plotting and to generate trees for filtered subsets of data combining molecular and clinical/socio-demographic information. Built on open source software, the full source code along with ready to use packages are available at http://www.evocell.org/inTB.

TB genotyping databases developed at Institut Pasteur de la Guadeloupe (IPG)

The first database was initiated more than fifteen years ago at Institut Pasteur de la Guadeloupe (IPG) when an undergraduate trainee named Jérôme Maïsetti took the initiative of entering the available spoligotype patterns from our own Caribbean isolates (n=218 strains) in an excel spreadsheet and pooled them with published data (n=392 isolates) from other countries. Once the patterns were sorted, we realized that one could not only define predominant patterns but also trace the origin of strains and their potential movements. This database of 610 spoligotypes was tentatively named SpolDB1, and led to the first ever description of 69 major spoligotype patterns in order to better understand TB origin and transmission (Sola *et al.,* 1999). Development of SpolDB1 was followed by the launch of SpolDB2 containing data on 3319 isolates (Sola *et al.,* 2001), and SpolDB3 on 13008 isolates grouped into 813 shared-types (containing 11,708 isolates) and 1300 orphan patterns (Filliol *et al.,* 2002; 2003). More recently, development of the fourth MySQL-based version SpolDB4 in 2006, n=39,295 clinical isolates (Brudey *et al.,* 2006), and SITVITWEB in 2012, n=62,582 clinical isolates (Demay *et al.,* 2012), permitted to have a finer phylogeographical snapshot of circulating MTBC genotypic lineages worldwide. The updated version of our database named SITVIT2 (n=111635 isolates) today contains genotyping information on approximately twice more strains than in the previous version, and will be released in 2014. It should be underlined that these two recent versions are MySQL-based multimarker databases with both spoligotyping and MIRU-VNTR data, limited to 12-loci MIRUs in SITVITWEB; and 12-, 15- or 24-loci MIRU-VNTR data in SITVIT2. In these recent versions, a web-based interface allows the user to search for strains through the database by criteria, such as the year,

Research

the isolation country, the country of origin, the investigator's name; as well as additional combined searches in SITVIT2, making it possible to get the genotyping data on selected strains in conjunction with their geographical distribution, as well as available data on drug-resistance, demographic and epidemiologic characteristics.

These successive developments of IPG databases have allowed to considerably improve our knowledge on genotyping and phylogeny/phylogeography of TB worldwide. The various versions of the our databases have led to a significant number of bilateral and multilateral studies as evidenced by the very high numbers of citations the successive databases have received (interrogation made on Google Scholar on March 31st 2014): SpolDB3 (Filliol *et al.,* 2002; 2003), 180 + 220 or 400 citations collectively; SpolDB4 (Brudey *et al.,* 2006), 661 citations; and SITVITWEB (Demay *et al.,* 2012); 67 citations, although made available only a year ago. We are therefore certain that the next version SITVIT2 will also find its place among the research community as a useful tool not only for TB molecular population genetics, historical demography and epidemiological modeling, but also for fundamental genetic analyses.

Brief description of SITVIT2

Future main functionalities of the SITVIT2 website will be improved as compared to the current SITVITWEB version both numerically as well as at the level of interface for future queries, although the lineage designations will be maintained almost unchanged with few exceptions. At the time of this study, SITVIT2 contained a total of 111,635 MTBC clinical isolates from 169 countries of patient origin. For data collection, we either enriched the database with genotyping results obtained at the institut Pasteur of Guadeloupe, or received from various co-investigators and collaborating laboratories, or those retrieved from published studies (Demay *et al.,* 2012). The website was developed using the java server pages (JSP) technology and embedded in a free Apache Tomcat application server (http://tomcat.apache.org), stored at the Institut Pasteur of Guadeloupe. The java technology was implemented as described earlier (Demay *et al.,* 2012). As previously, the description of the genetic characters of the clinical isolates accessible in SITVIT2 relies on a unique key identifier (IsoNumber) which summarizes information on the country of isolation, a laboratory code number, the year of isolation, a code for drug resistance information (0 to 4), and a unique isolate number given by the participating laboratory/hospital. In accordance with ethical guidelines concerning electronic treatment of data, this allows an anonymous number which can only be decoded by the data provider (the microbiological laboratory which supplied the data) to trace back their patient information but not by other users. SITVIT2 performs the automatized labelling system of SpolDB4 and SITVITWEB that attributes to each spoligotype present in 2 or more strains in the database a Spoligotype International Type (SIT) number, and to each MIRU profile present in 2 or more strains a MIRU International Type (MIT) number. The MIT numbers for 12-, 15- or 24- loci MIRU-VNTRs formats are called as 12-MIT, 15-MIT and 24-MIT while those restricted to 5 Exact Tandem Repeats are labeled as VIT (for VNTR International Type). Note that "orphan" designates patterns reported for a single isolate that does not correspond to any of the patterns recorded in the repository of the SITVIT2 database.

Some main website functionalities include various query tools such as spoligotype format conversion (binary to octal and vice-versa), data submission and analysis for spoligotyping and various MIRU formats, as well as criteria search for each marker individually or combined, year and country of isolation/origin, investigator's name, geographical distribution maps, and associated demographic, epidemiologic, and drug resistance information. It is thus possible to make individual queries or grouped queries by entering a duly formatted excel file (model available on website). For all queries, the user might expect a detailed report on markers (SIT and MIT numbers), phylogenetical lineages, and associated information available in the database in an anonymized format. One of the ongoing developments will include the possibility to look for correspondence of nomenclature of MIT patterns in SITVIT2 according to MIRU-VNTRplus nomenclature (note that a comparison of lineages between the 2 databases did not show major differences; results not shown).

Major phylogenetic lineages in SITVIT2

In SITVIT2, strains are classified in major phylogenetic clades assigned according to signatures provided earlier, which includes various MTBC members (AFRI, *M. africanum*; BOV, *M. bovis*; CANETTII, *M. canettii*; MICROTI, *M. microti*; PINI, M. pinnipedii), as well as for *M. tuberculosis* sensu stricto, i.e., the Beijing clade, the Central-Asian (CAS) clade, the East-African-Indian (EAI) clade, the Haarlem/Ural clades, the Latin-American-Mediterranean (LAM) clade, the Cameroon and Turkey lineages, the «Manu» family, the IS6110-low banding X clade, and the ill-defined T clade. Note that some spoligotypes previously classified as H3/H4 sublineages within Haarlem family were recently relabeled "Ural" (Mokrousov 2012); these include patterns belonging to H4 sublineage that were relabeled «Ural-2», and some patterns previously classified as H3 sublineage but with an additional specific signature (presence of spacer 2, absence of spacers 29 to 31, and 33 to 36), that are now relabeled «Ural-1». Furthermore, two LAM sublineages were recently raised to independent lineage level: LAM10-CAM as Cameroon lineage (Koro Koro *et al.,* 2013), and LAM7-TUR as Turkey lineage (Abadia *et al.,* 2010; Kisa *et al.,* 2012). We have kept this nomenclature unaltered for spoligotyping based genotypic lineages since it has already been found to be useful for local or global molecular epidemiological studies, as well as to follow the evolutionary and quantitative genetics of tubercle bacilli at a global scale. Note that the distribution of clinical isolates in SITVIT2 is studied both country wise and at macro-geographical level as sub-regions defined according to United Nations (according to http://unstats.un.org/unsd/methods/m49/m49regin.htm); 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) 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.

The readers are requested to refer to **Table 2** for a brief summary on the comparison of SITVITWEB vs. SITVIT2 and corresponding phylogenetical lineages in the 2 versions, and to **Figure 1** which highlights the evolution of strains recorded by various geographical sub-regions between the 2 versions. The most noticeable increase between the 2 versions can be observed for Southern Europe and Eastern Asia, followed

Research

by Central and Southern America, Northern, Eastern and Western Africa, and Northern and Western Europe (**Figure 1**). Lineage-wise, the proportion of Beijing genotype did not differ significantly between SITVITWEB and SITVIT2 (representing respectively 9.84% vs. 9.72% of isolates globally; **Table 2**); and was predominant in Asia, followed by significantly visible proportions in North America, South Africa, and Australasia. Proportions were also similar for CAS (3.69% vs. 3.91%), Cameroon (previously LAM10-CAM: 1.04% vs. 0.98%), Turkey (previously LAM7-TUR: 0.59% vs. 0.53%), and Manu (1.08% vs. 0.95%) lineages. However the most significant increase was observed for *M. bovis* strains (10.36% vs. 23.06%), underlining the increased potential of SITVIT2 not only to study M. *tuberculosis* epidemiology but also bovine tuberculosis.

Table 2. A summarized representation of the SITVITWEB and SITVIT2 databases and the corresponding major phylogenetic lineages of the M. tuberculosis complex (MTBC).

Major Lineages*	SITVITWEB * (n=62582)		SITVIT2 (n=111635)	
	Nb	%	Nb	%
Beijing	6,159	9.84	10,850	9.72
AFRI	695	1.11	965	0.86
BOV	6,486	10.36	25,741	23.06
CANETTII	12	0.02	12	0.01
CAS	2,480	3.96	4,362	3.91
EAI	4,674	7.47	6,617	5.93
Haarlem/Ural	7,058	11.28	10,580	9.48
LAM	8,042	12.85	12,245	10.97
Cameroon (previously LAM10-CAM)	650	1.04	1095	0.98
Turkey (previously LAM7-TUR)	370	0.59	593	0.53
Manu	675	1.08	1,064	0.95
MICROTI	29	0.05	29	0.03
PINI	152	0.24	159	0.14
S	1,151	1.84	1,606	1.44
T	12,038	19.24	17,947	16.08
X	4,088	6.53	4,683	4.19

Worldwide distribution maps in SITVIT2

The worldwide distribution map of major lineages in SITVIT2 illustrated in **Figure 2** highlights the global phylogeographical geo-specificities of MTBC isolates as seen in a 2014 snapshot. Even though these geo-specificities were already suggested in previous studies based both on spoligotyping (Brudey *et al.,* 2006; Demay *et al.,* 2012); and LSPs (Gagneux *et al.,* 2006), the present map corroborates the current distribution pattern and specificities thanks to curated data in SITVIT2 database. Furthermore, the reclassification of certain lineages that didn't appear in previous versions such as "Ural" shows that it is noticeably present in Russia, Central Asia, Southern Asia, and Western Asia, as well as in Finland (Northern Europe) which shares a common frontier and privileged links to Russia, specially as a buffer zone in a succession of wars between Russia and Sweden during the 18th century (http://www.historyworld.net/wrldhis/PlainTextHistories.asp?historyid=ad02). We further confirmed the relabeling of LAM7-TUR as Turkey lineage and that of LAM10-CAM as Cameroon lineage thanks to their phylogeographical distribution in SITVIT2 (**Figure 2**). It is important to underline that the Turkey lineage is progressing in Eastern European countries (representing around 6% of all MTBC strains in Eastern Europe in 2014 vs. less than 2% until 2006; results not shown). One may also notice the continuous reduction of the proportion of AFRI lineage in Western Africa, which represented around 37% of all MTBC strains in this sub-region until 2006 vs. 29% in 2014. This observation corroborates our previous suggestion that evolutionary ancestral *M. africanum* strains in Western Africa are slowly being replaced by evolutionary recent MTBC lineages such as Cameroon or other Euro-American lineages (Groenheit *et al.,* 2011). Last but not least, although the global proportions of CAS did not significantly change between SITVITWEB and SITVIT2 (3.69% vs. 3.91%), an increase was observed in Western Asia (12% vs. 18%) and East Africa (10% vs. 15%).

Looking for major associations between phylogenetic lineages vs. demographic and epidemiologic characteristics in SITVIT2

A user will be able to retrieve phylogenetic information linked to a number of different parameters such as incidence of the disease as seen through maps available from WHO, demography (age, sex-ratio) and other characteristics from various sub-regions in order to underline specificities in function of countries, regions, and populations. Considering that spreading of MDR- and XDR-TB clones in general populations represents a major threat in TB control, the interest of such a database is also to provide with a prevailing snapshot as well as to pinpoint any emergence for Public Health authorities. This geo-referencing of our data using Google API already provides with a potential Geographical Information Systems (GIS), that represents an efficient bacteriological counterpart of the WHO's Communicable Disease Global Atlas (http://apps.who.int/globalatlas/), allowing to bring together in a single electronic platform analysis and interpretation of genotyping data in conjunction with information on demography, socioeconomic conditions, and environmental factors. In our opinion, one should use such a mapping ideally in conjunction with suitable statistical and bioinformatical tools and softwares to better describe the TB genetic landscape. In addition to the tools and softwares described in previous studies (Brudey *et al.,* 2006; Demay *et al.,* 2012); we also use following tools in SITVIT2 analyses performed in routine:

(a) STATA software version 12 for descriptive and univariate analyses.

(b) R software version 2.14.1 to calculate the Odds Ratios (OR) and 95% Confidence Interval (CI) values.

(c) Pearson's Chi-square test and Fisher's Exact Test to compare major associations between genotyping data (shared types /

*The strains are classified in major phylogenetic clades assigned according to signatures provided earlier (Demay et al., 2012); which includes various MTBC members (AFRI, *M. africanum*; BOV, *M. bovis*; CANETTII, *M. canettii;* MICROTI, *M. microti*; PINI, *M. pinnipedii*), as well as for lineages/sub-lineages of *M. tuberculosis* sensu stricto (note that the sublineages are not shown), i.e., the Beijing clade, the Central-Asian (CAS) clade, the East-African-Indian (EAI) clade, the Haarlem/Ural clades,the Latin-American-Mediterranean (LAM) clade, the Cameroon and Turkey lineages, the «Manu» family, the IS6110-low banding X clade, and the ill-defined T clade.

Research

lineages) vs. demographic, epidemiologic, or socioeconomic characteristics (P values of <0.05 being considered as statistically significant).
(d) Minimum Spanning Trees (MSTs) are constructed based on genotyping data (spoligotypes, MIRUs) using BioNumerics software, version 6.6 (Applied Maths, Sint-Martens-Latem, Belgium). MSTs are connected undirected graphs in which all of the patterns are linked together with the fewest possible linkages between nearest neighbors.
(e) SpolTools software (http://www.emi.unsw.edu.au/spolTools) is used to draw Spoligoforests trees based on the Fruchterman-Reingold algorithm or a Hierarchical Layout (Reyes *et al.,* 2008; Tang *et al.,* 2008). Note that contrary to the MSTs, Spoligoforests are directed (and not necessarily connected) graphs allowing to highlight the evolutionary relationships between ascendant and descendant spoligotyping patterns.
(f) GraphViz software available at http://www.graphviz.org (Ellson *et al.,* 2002) to color the Spoligoforests in function of the lineages.
(g) WebLogo application version 2.8.2 (available at http://weblogo.berkeley.edu/(Schneider and Stephens, 1990; Crooks *et al.,* 2004) to evaluate and visualize the allelic diversity of the spoligotyping patterns in function of their associated lineages. This method of representation adapted to 43-spacer spoligotyping was labeled as "Spoligologos" (Driscoll *et al.,* 2002). WebLogos consist of stacks of symbols as graphical representations of the comparative diversity observed for individual spoligotyping spacer – one stack for each of the 43 spacer. The overall height of the stack indicates the conservation of a given spacer (the letter "n" designates the presence of a spacer, and letter "o" designates the absence). If a spacer is always the present or absent for one position of the 43 available ones (i.e., if 100% of the strains conserve the same presence or absence in one position, it corresponds to 4 bits), while the height of individual symbols within the stack indicates the relative frequency of absent/present spacer at that position.

Examples of some recent studies using SITVIT2 database

It is obviously not possible to detail all the studies done using SITVIT2, however some recent examples that looked for associations between phylogenetic lineages vs. demographic and epidemiologic parameters include:
(a) Geographical distribution map of spoligotyping-based MTBC lineages in various subregions of Africa and high phylogeographical specificity of *M. africanum* for Western Africa, with Guinea-Bissau being the epicenter (Groenheit *et al.,* 2011).
(b) Evidence that MTBC strains potentially involved in the TB epidemic in Sweden a century ago belonged to a closely knit pool of evolutionary recent PGG2/3 strains restricted to Sweden and its immediate neighbors (Groenheit *et al.,* 2012).
(c) Exploration of MTBC phylogenetic associations with drug resistance in Peru suggesting a prolonged, clonal, hospital-based outbreak of MDR disease amongst HIV patients (Sheen *et al.,* 2013).

(d) Phylogeographical mapping of TB in Finland showing a close resemblance of global MTBC population structure to one reported for Sweden, specially the predominance of the Euro-American family among elderly TB patients; the main difference being observed for the Ural lineage which was present in significant proportions among Finnish born cases (and also found in Russia, Latvia, and Estonia), but not in Sweden (Smit *et al.,* 2013).
(e) At the worldwide level using the SITVIT2 database, we find that X and LAM lineages are significantly more associated with HIV-positive serology; p-value<0.0001 (article in preparation).
(f) In several studies, the association of Beijing lineage with excessive drug resistance including MDR and/or XDR-TB was highlighted (van Soolingen *et al.,* 1995; Glynn *et al.,* 2002; Parwati *et al.,* 2010). Hence we recently performed exploration of phylogenetic associations (MTBC split in 2 groups as Beijing vs. other lineages) with drug resistance (quantified as pansusceptible, MDR-TB, XDR-TB, or any other drug resistance) using the SITVIT2 database (Couvin and Rastogi, 2014). The distribution of drug resistance for different subregions is shown in **Figure 3A**. Although proportion of drug-resistant strains was significantly higher for Beijing vs. non Beijing strains globally, important variations in the distribution of drug-resistance were observed. Drug resistance was significantly linked to Beijing (vs. non-Beijing strains) in Russia, Southern Asia, Southeastern Asia, and European countries, but not in Americas, Western Asia, China and Japan. If one considers the evolution of drug resistance over time for Beijing strains (1998 to 2011, **Figure 3B**), a continuous progression in the proportion of MDR and XDR strains (and a relative decrease of pansusceptible strains) is visible worldwide since 2003. Last but not least, we also observed that a rare but emerging spoligotype pattern (SIT190/Beijing) was significantly more associated with MDR-TB than the traditional SIT1/Beijing pattern (p-value<0.0001).
(g) Regarding the information provided thanks to a Spoligoforest tree, we would refer to a recent study conducted in Baghdad, Iraq (Mustafa Ali *et al.,* 2014). The results obtained on a total of 270 MTBC isolates showed that 2 specific patterns SIT1144/T1 and SIT309/CAS1-Delhi predominated in this study (6.3% for each pattern). The evolutionary relationships between Iraqi isolates as seen through a Spoligoforest tree using a Hierarchical Layout (**Figure 4**) clearly show that the bulk of TB in postwar Iraq is limited to 2 phylogenetically related group of MTBC strains belonging to T and CAS lineages.
(h) Regarding the information provided thanks to geo-referencing of genotyping data in SITVIT2 using Google API, we could refer to the same study (Mustafa Ali *et al.,* 2014). As shown in **Figure 5**, significant differences were highlighted between Baghdad city as compared to other cities in Iraq, regarding both demographics and drug resistance information. Indeed, with a male/female sex ratio of 1.49 in Baghdad vs. 3.04 in other governorates, the proportion of female patients was significantly higher in Baghdad city (p-value = 0.009; Odds Ratio = 0.49 and 95%CI [0.28; 0.86]). The proportion of newly treated vs. re-treated cases differed significantly between the 2 groups, the proportion of re-treated patients being higher in other governorates of Iraq (p-value=0.007; OR=2.1, 95%CI [1.19; 3.62]). Finally, the rate of MDR-TB was higher in other governorates than Baghdad (p >0.12; difference not statistically significant).

Conclusions

The collection of databases developed at Institut Pasteur de la

Research

Figure 1. (A) Worldwide distribution of MTBC isolates recorded in both SITVITWEB (blue circle) and SITVIT2 (green circle), the number of isolates per sub-region is indicated inside each circle. (B). Gradual evolution of series of databases (from SpolDB2 to SITVIT2) in function of the number of MTBC isolates for each database.

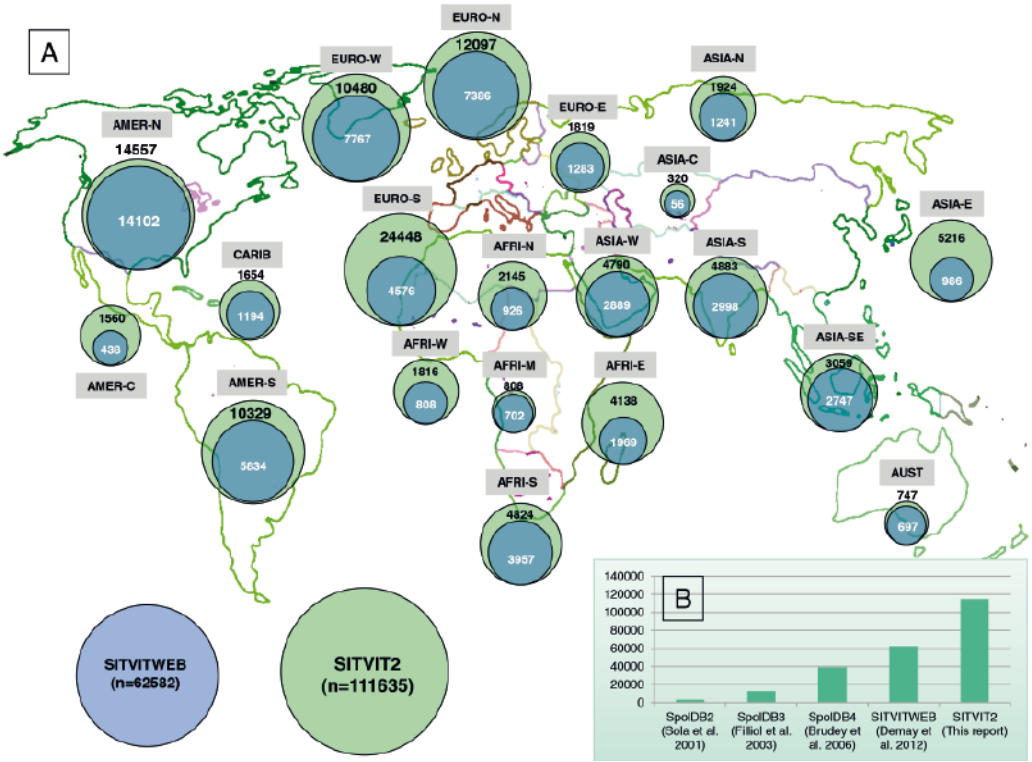
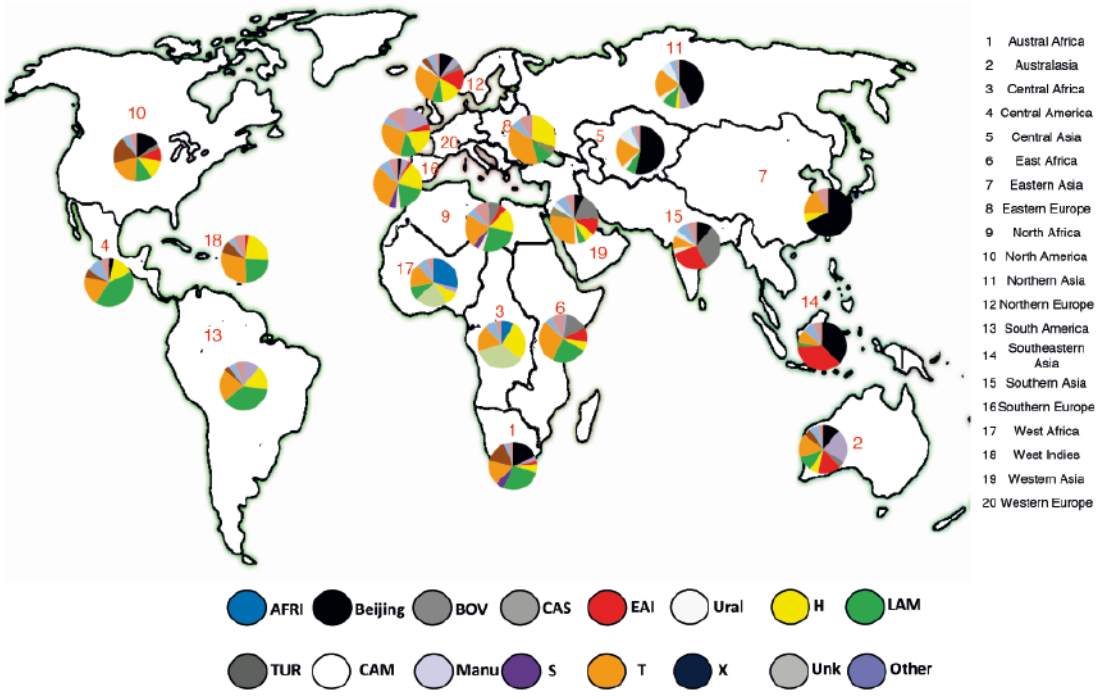


Figure 2 : Worldwide distribution of lineages contained in the SITVIT2 database.



Research

Guadeloupe has helped achieve a global overview of worldwide TB situation. With the newly updated SITVIT2 database, we have been further able to highlight the ongoing circulation of MTBC strains based on extended genotyping markers, as well as underlined major associations between MTBC phylogenetic lineages vs. demographic and epidemiologic characteristics. Future developments should ideally include inclusion of other markers such as RD/LSP and SNPs, as well as the future information that is going to be generated thanks to next generation sequencing. Although many evolutionary and pathobiological characteristics of the prevailing TB epidemic remain to be discovered, the new SITVIT collection of databases represents a major tool for an improved epidemiological surveillance and control of TB.

Acknowledgements

We thank more than 500 investigators who provided data for SpolDB4, SITVITWEB, and SITVIT2 databases (among which 190 contributed genotyping data on more than 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 Thierry Zozio, Julie Millet, Veronique Hill, and Elisabeth Streit for helpful discussions. DC was awarded a Ph.D. fellowship by the European Social Funds through the Regional Council of Guadeloupe.

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*. 2010; 107: 1066–1074.

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.

Allix-Béguec C, Wahl C, Hanekom M, Nikolayevskyy V, Drobniewski F, Maeda S, Campos-Herrero I, Mokrousov I, Niemann S, Kontsevaya I, Rastogi N, Samper S, Sng LH, Warren RM, Supply P. 2014. Proposal of a consensus set of hypervariable mycobacterial interspersed repetitive-unit-variable-number tandem-repeat loci for subtyping of *Mycobacterium tuberculosis* Beijing isolates. *J Clin Microbiol*. 52: 164-172.

Barry CE 3rd, Boshoff HI, Dartois V, Dick T, Ehrt S, Flynn J, Schnappinger D, Wilkinson RJ, Young D. 2009. The spectrum of latent tuberculosis: rethinking the biology and intervention strategies. *Nat Rev Microbiol*. 7: 845-855.

de Boer AS, Borgdorff MW, de Haas PE, Nagelkerke NJ, van Embden JD, van Soolingen D. 1999. Analysis of rate of change of IS6110 RFLP patterns of *Mycobacterium tuberculosis* based on serial patient isolates. *J Infect Dis*. 180: 1238-1244.

Brosch R, Gordon SV, Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, Garnier T, Gutierrez C, Hewinson G, Kremer K, Parsons LM, Pym AS, Samper S, van Soolingen D, Cole ST. 2002. A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proc Natl Acad Sci USA*. 99: 3684-3689.

Brudey K, Driscoll JR, Rigouts L, Prodinger WM, Gori A, Al-Hajoj SA, Allix C, Aristimuño L, Arora J, Baumanis V, Binder L, Cafrune P, Cataldi A, Cheong S, Diel R, Ellermeier C, Evans JT, Fauville-Dufaux M, Ferdinand S, Garcia de Viedma D, Garzelli C, Gazzola L, Gomes HM, Guttierez MC, Hawkey PM, van Helden PD, Kadival GV, Kreiswirth BN, Kremer K, Kubin M, Kulkarni SP, Liens B, Lillebaek T, Ho ML, Martin C, Martin C, Mokrousov I, Narvskaïa O, Ngeow YF, Naumann L, Niemann S, Parwati I, Rahim Z, Rasolofo-Razanamparany V, Rasolonavalona T, Rossetti ML, Rüsch-Gerdes S, Sajduda A, Samper S, Shemyakin IG, Singh UB, Somoskovi A, Skuce RA, van Soolingen D, Streicher EM, Suffys PN, Tortoli E, Tracevska T, Vincent V, Victor TC, Warren RM, Yap SF, 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.

CDC. 2010. Launch of TB Genotyping Information Management System (TB GIMS). *Morbidity and Mortality Weekly Report* (MMWR) March 19, 59(10); 300.

Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, Parkhill J, Malla B, Berg S, Thwaites G, Yeboah-Manu D, Bothamley G, Mei J, Wei L, Bentley S, Harris SR, 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.

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.

Couvin D, Rastogi N. (2014). Tuberculosis – a global emergency: tools and methods to monitor, understand, and control the epidemic with specific example of the Beijing lineage. Tuberculosis (Edinb). Submitted for publication.

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res*. 14: 1188-1190.

Dale JW, Brittain D, Cataldi AA, Cousins D, Crawford JT, Driscoll J, Heersma H, Lillebaek T, Quitugua T, Rastogi N, Skuce RA, Sola C, Van Soolingen D, Vincent V. 2001. Spacer oligonucleotide typing of bacteria of the *Mycobacterium tuberculosis* complex: recommendations for standardised nomenclature. *Int J Tuberc Lung* Dis. 5: 216-219.

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.

Driscoll JR, Bifani PJ, Mathema B, McGarry MA, Zickas GM, Kreiswirth BN, Taber HW. 2002. Spoligologos: a bioinformatic approach to displaying and analyzing *Mycobacterium tuberculosis* data. *Emerg Infect Dis*. 8(11): 1306-9.

Ellson J, Gansner E, Koutsofios L, North SC, Woodhull G. 2002. Graphviz – Open Source Graph Drawing Tools. In: Mutzel P, Jünger M, Leipert S (Editors), Heidelberg: Springer-Verlag Berlin. pp. 483-484.

van Embden JDA, Cave MD, Crawford JT, Dale JW, Eisenach KD, Gicquel B, Hermans P, Martin C, McAdam R, Shinnick TM, Small PM. 1993. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. *J Clin Microbiol*. 31, 406-409.

Fang Z, Kenna DT, Doig C, Smittipat DN, Palittapongarnpim P, Watt B, Forbes KJ. 2001. Molecular evidence for independent occurrence of IS6110 insertions at the same sites of the genome of *Mycobacterium tuberculosis* in different clinical isolates. *J Bacteriol*. 183: 5279-5284.

Fenner L, Malla B, Ninet B, Dubuis O, Stucki D, Borrell S, Huna T, Bodmer T, Egger M, Gagneux S. (2011) "Pseudo-Beijing": evidence for convergent evolution in the direct repeat region of *Mycobacterium tuberculosis*. *PLoS One*. 6: e24737.

Filliol I, Motiwala AS, Cavatore M, Qi W, Hazbón MH, Bobadilla del Valle M, Fyfe J, García-García L, Rastogi N, Sola C, Zozio T, Guerrero MI, León CI, Crabtree J, Angiuoli S, Eisenach KD, Durmaz R, Joloba ML, Rendón A, Sifuentes-Osornio J, Ponce de León A, Cave MD, Fleischmann R, Whittam TS, 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*. 2006; 188: 759-772.

Filliol I, Driscoll JR, Van Soolingen D, Kreiswirth BN, Kremer K, Valétudie G, Anh DD, Barlow R, Banerjee D, Bifani PJ, Brudey K, Cataldi A, Cooksey RC, Cousins DV, Dale JW, Dellagostin OA, Drobniewski F, Engelmann G, Ferdinand S, Gascoyne-Binzi D, Gordon M, Gutierrez

Research

Figure 3. Drug resistance characteristics of Beijing vs. non-Beijing *M. tuberculosis* lineages (A) and evolution of drug resistance among Beijing isolates between 1998 to 2011 (B).

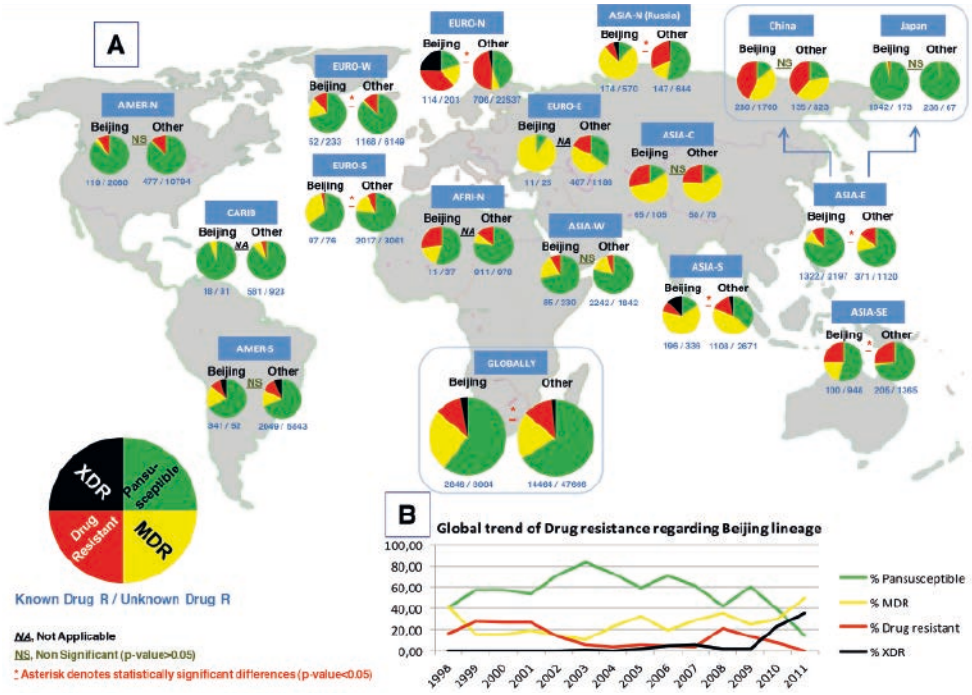
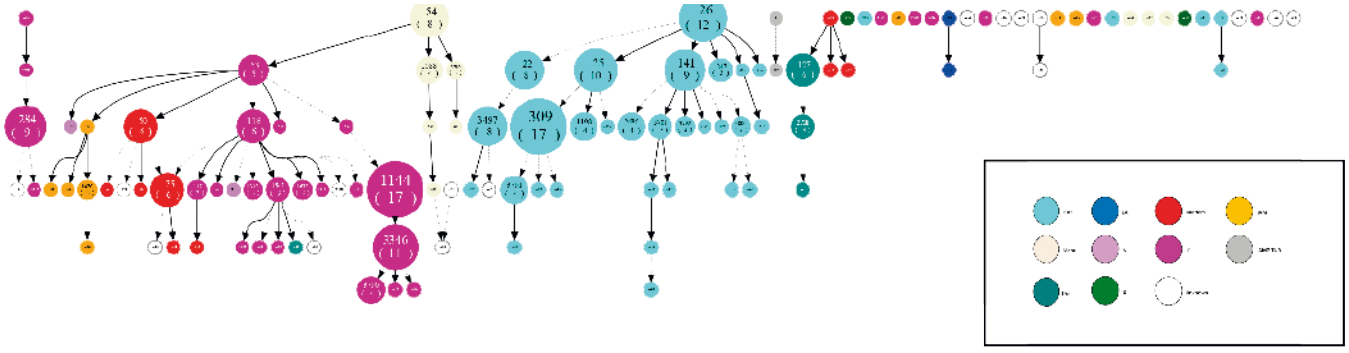
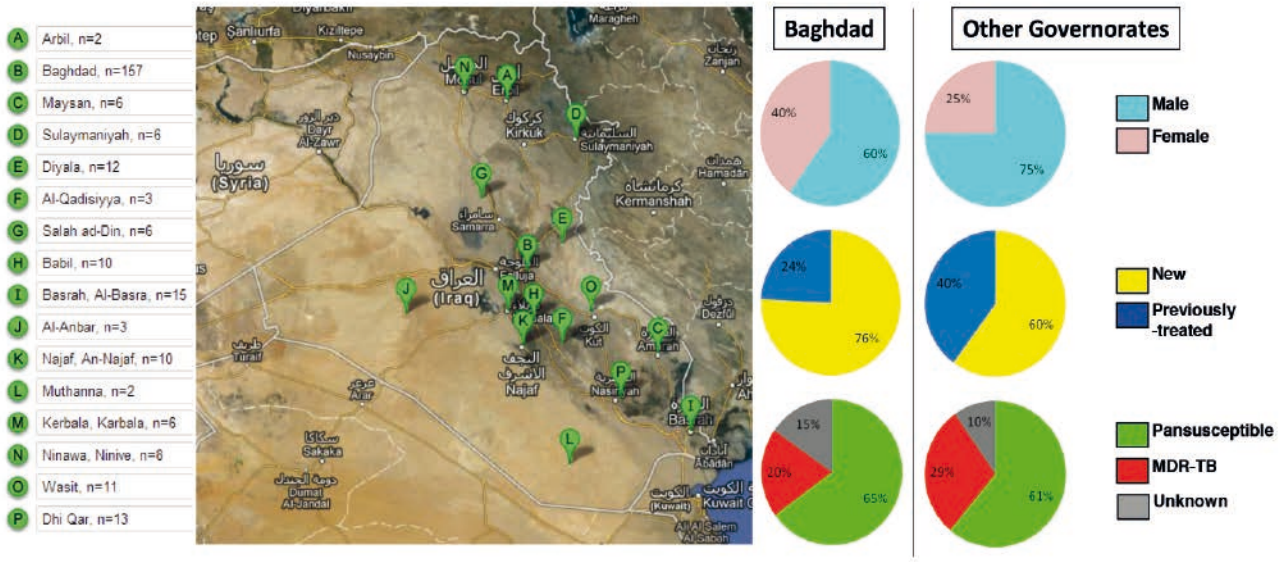


Figure 4. A Spoligoforest representation of parent to descendant spoligotypes in a study from Baghdad, Iraq (n=270 isolates) using a Hierarchical Layout. In this tree, each spoligotype pattern is represented by a node with area size being proportional to the total number of isolates with that specific pattern. Changes (loss of spacers) are represented by directed edges between nodes, with the arrowheads pointing to descendant spoligotypes. The heuristic used selects a single inbound edge with a maximum weight using a Zipf model. Solid black lines link patterns that are very similar, i.e., loss of one spacer only (maximum weigh being 1.0), while dashed lines represent links of weight comprised between 0.5 and 1, and dotted lines a weight less than 0.5. Note that SIT309/CAS1-Delhi and SIT1144/T1 are the biggest nodes (n=17), followed by SIT26/CAS1-Delhi (n=12), SIT3346/T1 (n=11) and SIT25/CAS1-Delhi (n=10), which are other predominant patterns in our study. Finally, orphan isolates (double circled), appear mostly at terminal positions on the tree, or are isolated strains without interconnections with the other strains (figure based on data from reference 70).



Research

Figure 5. A Google Map screenshot showing MTBC strain distribution in function of city of isolation in Baghdad vs. other cities in Iraq (n=270 isolates) and comparison of data in function of patient origin vs. sex-ratio, treatment status, and drug resistance for Baghdad vs. other governorates (figure based on data from reference 70).



MC, Haas WH, Heersma H, Källenius G, Kassa-Kelembho E, Koivula T, Ly HM, Makristathis A, Mammina C, Martin G, Moström P, Mokrousov I, Narbonne V, Narvskaya O, Nastasi A, Niobe-Eyangoh SN, Pape JW, Rasolofo-Razanamparany V, Ridell M, Rossetti ML, Stauffer F, Suffys PN, Takiff H, Texier-Maugein J, Vincent V, De Waard JH, Sola C, Rastogi N. 2002. Global distribution of *Mycobacterium tuberculosis* spoligotypes. *Emerg Infect Dis*. 2002 Nov; 8(11): 1347-9.

Filliol I, Driscoll JR, van Soolingen D, Kreiswirth BN, Kremer K, Valétudie G, Dang DA, Barlow R, Banerjee D, Bifani PJ, Brudey K, Cataldi A, Cooksey RC, Cousins DV, Dale JW, Dellagostin OA, Drobniewski F, Engelmann G, Ferdinand S, Gascoyne-Binzi D, Gordon M, Gutierrez MC, Haas WH, Heersma H, Kassa-Kelembho E, Ho ML, Makristathis A, Mammina C, Martin G, Moström P, Mokrousov I, Narbonne V, Narvskaya O, Nastasi A, Niobe-Eyangoh SN, Pape JW, Rasolofo-Razanamparany V, Ridell M, Rossetti ML, Stauffer F, Suffys PN, Takiff H, Texier-Maugein J, Vincent V, de Waard JH, 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(5): 1963-1970.

Gagneux S, DeRiemer K, Van T, Kato-Maeda M, de Jong BC, Narayanan S, Nicol M, Niemann S, Kremer K, Gutierrez MC, Hilty M, Hopewell PC, Small PM. 2006. Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *Proc Natl Acad Sci USA*. 103(8): 2869-73.

García de Viedma D, Mokrousov I, Rastogi N. 2011. Innovations in the molecular epidemiology of tuberculosis. *Enferm Infecc Microbiol* Clin. 29 (Suppl 1): 8-13.

Glynn JR, Whiteley J, Bifani PJ, Kremer K, van Soolingen D. (2002). Worldwide occurrence of Beijing/W strains of *Mycobacterium tuberculosis*: a systematic review. *Emerg Infect Dis*. 8(8): 843-849.

Gordon SV, Brosch R, Billault A, Garnier T, Eiglmeier K, Cole ST. 1999. Identification of variable regions in the genomes of tubercle bacilli using bacterial artificial chromosome arrays. *Mol Microbiol*. 32: 643-655.

Groenheit R, Ghebremichael S, Svensson J, Rabna P, Colombatti R, Riccardi F, Couvin D, Hill V, Rastogi N, Koivula T, Källenius G. 2011. The Guinea-Bissau family of *Mycobacterium tuberculosis* complex revisited. *PLoS One*. 6(4): e18601.

Groenheit R, Ghebremichael S, Pennhag A, Jonsson J, Hoffner S, Couvin D, Koivula T, Rastogi N, Källenius G. 2012. *Mycobacterium tuberculosis* strains potentially involved in the TB epidemic in Sweden a century ago. *PLoS One*. 7(10): e46848.

Gutacker MM, Mathema B, Soini H, Shashkina E, Kreiswirth BN, Graviss EA, Musser JM. Single-nucleotide polymorphism-based population genetic analysis of *Mycobacterium tuberculosis* strains from 4 geographic sites. *J Infect Dis*. 2006; 193: 121-128.

Heersma HF, Kremer K, van Embden JD. 1998. Computer analysis of IS6110 RFLP patterns of *Mycobacterium tuberculosis*. *Methods Mol Biol*. 101: 395-422.

Hermans PW, van Soolingen D, Bik EM, de Haas PE, Dale JW, van Embden JD. 1991. Insertion element IS987 from *Mycobacterium bovis* BCG is located in a hot-spot integration region for insertion elements in *Mycobacterium tuberculosis* complex strains. *Infect Immun*. 59: 2695-2705.

Hershberg R, Lipatov M, Small PM, Sheffer H, Niemann S, Homolka S, Roach JC, Kremer K, Petrov DA, Feldman MW, Gagneux S. 2008. High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. *PLoS Biol*. 2008, 6(12): e311.

Hirsh AE, Tsolaki AG, DeRiemer K, Feldman MW, Small PM. Stable association between strains of *Mycobacterium tuberculosis* and their human host populations. *Proc Natl Acad Sci USA*. 2004; 101: 4871-4876.

Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, Hill DP, Kania R, Schaeffer M, St Pierre S, Twigger S, White O, Rhee SY. (2008) Big data: The future of biocuration. *Nature*. 455(7209): 47-50.

Jagielski T, van Ingen J, Rastogi N, Dziadek J, Mazur PK, Bielecki J. 2014. Current Methods in the Molecular Typing of *Mycobacterium tuberculosis* and Other Mycobacteria. *Biomed Res Int*. 2014 (Article ID 645802): 1-21.

Joshi KR, Dhiman H, Scaria V. (2013) tbvar: a comprehensive genome variation resource for *Mycobacterium tuberculosis*. Database Vol. 2013: article ID bat083.

Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper S, Bunschoten A, Molhuizen H, Shaw R, Goyal M, van

Research

Embden J. 1997. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *J Clin Microbiol*. 35: 907-914.

Kato-Maeda M, Rhee JT, Gingeras TR, Salamon H, Drenkow J, Smittipat N, Small PM. 2001. Comparing genomes within the species *Mycobacterium tuberculosis*. *Genome Res*. 11: 547-554.

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*. 2012; 7(1): e30331.

Koro Koro F, Kamdem Simo Y, Piam FF, Noeske J, Gutierrez C, Kuaban C, Eyangoh SI. 2013. Population dynamics of *tuberculous Bacilli* in Cameroon as assessed by spoligotyping. *J Clin Microbiol.* 51: 299-302.

Mahairas GG, Sabo PJ, Hickey MJ, Singh DC, Stover CK. 1996. Molecular analysis of genetic differences between *Mycobacterium bovis* BCG and virulent M. bovis. J Bacteriol. 178: 1274-1282.

Mokrousov I. 2012. The quiet and controversial: Ural family of *Mycobacterium tuberculosis*. *Infect Genet Evol*. 12: 619-629.

Mustafa Ali R, Trovato A, Couvin D, Al-Thwani AN, Borroni E, Dhaer FH, Rastogi N, Cirillo DM. 2014. Molecular Epidemiology and Genotyping of *Mycobacterium tuberculosis* Isolated in Baghdad. BioMed Research International, vol. 2014, Article ID 580981, 15 pages.

Parwati I, van Crevel R, van Soolingen D. 2010. Possible underlying mechanisms for successful emergence of the *Mycobacterium tuberculosis* Beijing genotype strains. *Lancet Infect Dis*. 10: 103-111.

Radhakrishnan I, K MY, Kumar RA, Mundayoor S. 2001. Implications of low frequency of IS6110 in fingerprinting field isolates of *Mycobacterium tuberculosis* from Kerala, India. *J Clin Microbiol*. 39: 1683.

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

Reyes JF, Francis AR, Tanaka MM. 2008. Models of deletion for visualizing bacterial variation: an application to tuberculosis spoligotypes. *BMC Bioinformatics*. 9: 496.

Rodriguez-Campos S, González S, de Juan L, Romero B, Bezos J, Casal C, Álvarez J, Fernández-de-Mera IG, Castellanos E, Mateos A, Sáez-Llorente JL, 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(4): 877-82.

Rothschild BM, Martin LD, Lev G, Bercovier H, Bar-Gal GK, 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.

Salamon H, Segal MR, Ponce de Leon A, Small PM. 1998. Accommodating error analysis in comparison and clustering of molecular fingerprints. *Emerg Infect Dis*. 4: 159-168.

Schneider TD, Stephens RM. 1990. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 18: 6097-6100.

Shabbeer A, Cowan LS, Ozcaglar C, Rastogi N, Vandenberg SL, Yener B, Bennett KP. 2012. TB-Lineage: an online tool for classification and analysis of strains of *Mycobacterium tuberculosis* complex. *Infect Genet Evol.* 12: 789-797.

Sheen P, Couvin D, Grandjean L, Zimic M, Dominguez M, Luna G, Gilman RH, Rastogi N, Moore DA. 2013. Genetic diversity of *Mycobacterium tuberculosis* in Peru and exploration of phylogenetic associations with drug resistance. *PLoS One*. 8(6): e65873.

Smit PW, Haanperä M, Rantala P, Couvin D, Lyytikäinen O, Rastogi N, Ruutu P, Soini H. (2013). Molecular epidemiology of tuberculosis in Finland, 2008-2011. *PLoS One*. 8(12): e85027.

Smith NH, Upton P. 2012. Naming spoligotype patterns for the RD9-deleted lineage of the *Mycobacterium tuberculosis* complex; www.Mbovis.org. *Infect Genet Evol*. 12 (4), pp. 873–876.

Soares P, Alves RJ, Abecasis AB, Penha-Gonçalves C, Gomes MG, Pereira-Leal JB. (2013) inTB – a data integration platform for molecular and clinical epidemiological analysis of tuberculosis. *BMC Bioinformatics*. 14: 264.

Sola C, Devallois A, Horgen L, Maïsetti J, Filliol I, Legrand E, Rastogi N. 1999. Tuberculosis in the Caribbean: using spacer oligonucleotide typing to understand strain origin and transmission. *Emerg Infect Dis*. 5(3): 404-14.

Sola C, Filliol I, Gutierrez MC, 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(3): 390-396.

Sola C, Filliol I, Legrand E, Mokrousov I, Rastogi N. 2001. *Mycobacterium tuberculosis* phylogeny reconstruction based on combined numerical analysis with IS1081, IS6110, VNTR, and DR-based spoligotyping suggests the existence of two new phylogeographical clades. *J Mol Evol*. 53: 680-689.

Sola C, Filliol I, Legrand E, Lesjean S, Locht C, Supply P, Rastogi N. 2003. Genotyping of the *Mycobacterium tuberculosis* complex using MIRUs: association with VNTR and spoligotyping for molecular epidemiology and evolutionary genetics. *Infect Genet Evol*. 3: 125-133.

van Soolingen D, Qian L, de Haas PE, Douglas JT, Traore H, Portaels F, Qing HZ, Enkhsaikan D, Nymadawa P, van Embden JD. (1995). Predominance of a single genotype of *Mycobacterium tuberculosis* in countries of East Asia. *J Clin Microbiol.* 33: 3234-3238.

Sreevatsan S, Pan X, Stockbauer KE, Connell ND, Kreiswirth BN, Whittam TS, Musser JM. 1997. Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. *Proc Natl Acad Sci USA*. 94: 9869-9874.

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

Supply P, Allix C, Lesjean S, Cardoso-Oelemann M, Rüsch-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 MC, 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.

Tang C, Reyes JF, Luciani F, Francis AR, Tanaka MM. 2008. SpolTools: online utilities for analyzing spoligotypes of the *Mycobacterium tuberculosis* complex. *Bioinformatics*. 24: 2414-415.

Wolfe ND, Dunavan CP, Diamond J. 2007. Origins of major human infectious diseases. *Nature*. 447: 279-283.

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

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(Web Server issue): W326–W331.

69.WHO. 2006. The Stop TB Strategy: Building on and enhancing DOTS to meet the TB-related Millennium Development Goals. http://whqlibdoc.who.int/hq/2006/WHO_HTM_STB_2006.368_eng.pdf?ua=1.

WHO. 2013. Global tuberculosis report 2013.

www.who.int/iris/bitstream/10665/91355/1/9789241564656_eng.pdf