Tuberculosis 105 (2017) 49–52

Contents lists available at ScienceDirect

Tuberculosis

journal homepage: http://intl.elsevierhealth.com/journals/tube

SpolSimilaritySearch – A web tool to compare and search similarities between spoligotypes of *Mycobacterium tuberculosis* complex

David Couvin**, Thierry Zozio, Nalin Rastogi*

WHO Supranational TB Reference Laboratory, Tuberculosis & Mycobacteria Unit, Institut Pasteur de la Guadeloupe, Abymes, Guadeloupe, France

ARTICLE INFO

Article history:
Received 14 December 2016
Accepted 19 April 2017

Keywords:
Mycobacterium tuberculosis
Spoligotyping
Database
Molecular epidemiology
Evolution
Similarity search

ABSTRACT

Spoligotyping is one of the most commonly used polymerase chain reaction (PCR)-based methods for identification and study of genetic diversity of *Mycobacterium tuberculosis* complex (MTBC). Despite its known limitations if used alone, the methodology is particularly useful when used in combination with other methods such as mycobacterial interspersed repetitive units – variable number of tandem DNA repeats (MIRU-VNTRs). At a worldwide scale, spoligotyping has allowed identification of information on 103,856 MTBC isolates (corresponding to 98049 clustered strains plus 5807 unique isolates from 169 countries of patient origin) contained within the SITVIT2 proprietary database of the Institut Pasteur de la Guadeloupe. The SpolSimilaritySearch web-tool described herein (available at: http://www.pasteur-guadeloupe.fr:8081/SpolSimilaritySearch) incorporates a similarity search algorithm allowing users to get a complete overview of similar spoligotype patterns (with information on presence or absence of 43 spacers) in the aforementioned worldwide database. This tool allows one to analyze spread and evolutionary patterns of MTBC by comparing similar spoligotype patterns, to distinguish between widespread, specific and/or confined patterns, as well as to pinpoint patterns with large deleted blocks, which play an intriguing role in the genetic epidemiology of *M. tuberculosis*. Finally, the SpolSimilaritySearch tool also provides with the country distribution patterns for each queried spoligotype.

1. Introduction

Spacer oligonucleotide typing or Spoligotyping [1,2] is a polymerase chain reaction (PCR)-based hybridization assay that allows to detect the variability in the direct repeat (DR) region of *Mycobacterium tuberculosis* complex (MTBC). Since the DR region consists of multiple copies of a conserved 36-base-pair sequence (the direct repeats) separated by multiple unique spacer sequences (the standard assay uses 43 spacers); a positive or negative hybridization for each spacer theoretically allows for a total of 2^{43} (2 exponent 43) combination possibilities (i.e., 8796,093,022,208 possible patterns). Different *M. tuberculosis* strains are characterized by various complements of the 43 spacers, which has led to the widespread use of this method to study the *M. tuberculosis* molecular epidemiology [3]. Detection and identification of similar/related patterns along with confinement/rarity of certain patterns is important both for understanding and evaluating probable routes of tuberculosis (TB) transmission as well as for a better comprehension of *M. tuberculosis* evolutionary aspects. Despite the fact that spoligotyping used alone suffers from some limitations, it allows for a better and rapid discrimination of *M. tuberculosis* isolates when used in combination with other methods such as mycobacterial interspersed repetitive units – variable number of tandem DNA repeats (MIRU-VNTRs) [3–5].

Several databases and web based applications were developed using spoligotyping-based genotyping data [6–9]. More recently, online tools allowing to infer in silico spoligotyping patterns from next-generation sequencing reads have also been made available [10,11]. We consequently developed a tool, SpolSimilaritySearch, which allows a better visualization of similar spoligotyping patterns having a particular signature, as well as their worldwide country-based distribution. SpolSimilaritySearch fosters analyzing spread and evolutionary history of MTBC strains, and is able to highlight shared patterns vs. specific and/or confined patterns.

* Corresponding author. WHO Supranational TB Reference Laboratory, Tuberculosis & Mycobacteria Unit, Institut Pasteur de la Guadeloupe, Morne Jolivière, BP 484, 97183 Abymes, Guadeloupe, France.

** Corresponding author. WHO Supranational TB Reference Laboratory, Tuberculosis & Mycobacteria Unit, Institut Pasteur de la Guadeloupe, Morne Jolivière, BP 484, 97183 Abymes, Guadeloupe, France.

E-mail addresses: david.couvin@googlemail.com (D. Couvin), tzozio@pasteur-guadeloupe.fr (T. Zozio), nrastogi@pasteur-guadeloupe.fr (N. Rastogi).

http://dx.doi.org/10.1016/j.tube.2017.04.007
1472-9792/

Besides, it enables to pinpoint spoligotyping patterns with large deleted blocks, which plays an intriguing yet not fully understood role in the genetic epidemiology of *M. tuberculosis* [12]

2. Materials and methods

SpolSimilaritySearch was developed as a publicly accessible web tool (http://www.pasteur-guadeloupe.fr:8081/SpolSimilaritySearch) to enable viewing and comparison of similar spoligotyping patterns. Spoligotyping data on 103,856 isolates (corresponding to 98049 clustered strains plus 5807 unique isolates from 169 countries of patient origin) exploited through the SpolSimilaritySearch tool were extracted from the proprietary databases of the Institut Pasteur de la Guadeloupe – SITVITWEB [7] and SITVIT2 [9].

Data on 43 spacers were duly checked for each strain and curated, anonymized and used without patient identification. Country distribution was provided for each pattern using two-letter country codes (ISO 3166-1 alpha-2; http://en.wikipedia.org/wiki/ISO_3166-1_alpha-2), and further classified within a defined MTBC phylogenetical lineage. Lastly, each pattern shared by 2 or more patients was assigned a Spoligotype International Type (SIT) number. The web interface was implemented using Java technology (Java Server Pages, Asynchronous JavaScript, Ajax), Google Code API, and XML under the integrated development environment (IDE) Eclipse (https://www.eclipse.org/). Data were integrated within a MySQL Server database. In addition to classical SQL queries, the search of specific binary spoligotyping signatures by regular expressions (http://en.wikipedia.org/wiki/Regular_expression) allowed an improved interrogation. For examining similarity, a simple search algorithm is used through Java Server Pages and MySQL database. The spoligotype binary pattern is provided by selecting presence, absence or uncertainty (presence or absence) of 43 spacers; the binary profile is then translated into octal spoligotype (including uncertainty coded using regular expressions), and finally, all patterns matching with the given spoligotype are retrieved from the SITVIT2 database with their characteristics. This web application is hosted and deployed on an Apache Tomcat Server (version 6) at the Institut Pasteur de la Guadeloupe. Note that in response to a query, only limited anonymized data is provided by the web-tool, which we have provided as an "OCR-compatible" pdf file (Supplemental File S1). For each strain, we have also provided the genotyping results both in "Octal" as well as "Binary" formats, which are interconvertible using the online spoligotype conversion tool available at: http://www.pasteur-guadeloupe.fr:8081/SITVIT_ONLINE/tools.jsp#

3. Results and discussion

As detailed in Supplementary File S1, a total of 3851 SITs (corresponding to 98049 isolates) were identified, bringing the total number of distinct patterns to 3851 shared + 5807 unique patterns or 9658 spoligotyping patterns. For each pattern, the country distribution was also provided. Among these patterns, n = 8033 were assigned to a phylogenetical lineage, while n = 1625 patterns remained unclassified (with an unknown signature). The major lineages were assigned according to signatures provided earlier [3,7,9]; which included 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, 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 IS*6110*-low banding X clade, and the ill-defined T clade. Information on numbers and percentages of various lineages is provided elsewhere [3,9].

As summarized recently [3], this global MTBC population structure defined by distinct lineages, is corroborated by different genetic markers; nevertheless, the lineage nomenclature significantly varies depending on the marker used. For example, the EAI lineage denotes two completely different MTBC phylogenetical groups by spoligotyping versus Large Sequence Polymorphisms (LSP). Thus the spoligotyping-based "ancestral EAI lineage" corresponds to LSP-based "Indo-Oceanic" lineage, while the LSP-based EAI corresponds to spoligotyping-based "CAS" lineage [3]. To avoid any confusion, the readers are referred to a recent review article for a detailed comparison of nomenclature of MTBC lineages by spoligotyping versus other markers, i.e., *katG-gyrA* principal genetic groupings (PGG), SNP-based clusters groups (SCG), and Large Sequence Polymorphisms (LSP)/Regions of Difference (RD) based lineages [3].

A major contribution of our global repository – housed at Institut Pasteur de la Guadeloupe – is the simplicity with which one can map spoligotyping data to define the TB genetic landscape. To further streamline the process of comparison and similarity search between numerous MTBC spoligotypes, we decided to develop a dedicated web tool so researchers could easily compare their own spoligotyping patterns in function of the presence, absence, or uncertainty (presence or absence) of spacers within a 43 spacer format, against a collection of 103,856 MTBC isolates from 169 countries of patient origin, contained within the SITVIT2 proprietary database (Supplementary Table S1). This huge collection of spoligotyping patterns allows users to proceed to a kind of alignment of similar profiles, allowing to decipher their phylogeographical distribution and specificity, and eventually to underline the geographical rarity and/or confinement of certain spoligotypes.

The web interface of SpolSimilaritySearch (Fig. 1) was designed to be user-friendly, allowing to search and visualize spoligotyping patterns in function of the presence, absence, or uncertainty (presence or absence) of spacers within a 43 spacer format. As detailed in the user manual (Supplementary File S2), individual users have the possibility to select each spacer as being: (i) present, by selecting the black square (■); (ii) absent, by selecting the white square (□); and (iii) uncertain or variable, by selecting the dash symbol (–); a functionality which allows a true representation of all the 43 different spacers. Submitting their queries by a simple click, users immediately obtain a detailed list of spoligotyping patterns followed by their SIT number, lineage, and country distribution in the SITVIT2 database. Queries on variable presence or absence of a given spacer return all possibilities considering the various topographies allowing a user to double-check their data on the basis of lineage attribution and country distribution profiles.

SpolSimilaritySearch tool was successfully used recently in studies focusing on spoligotyping-based MTBC population structure [13,14]. Some examples of usefulness of SpolSimilaritySearch tool are provided in Supplementary File S3. For example, in a recent study by Ismail et al., 2014 [13] focusing on MTBC genotypic diversity in Malaysia, we investigated if a predominant spoligotype pattern corresponding to the ancestral EAI lineage could be linked to a Malaysia-specific signature. SpolSimilaritySearch indeed allowed to highlight the specificity of spoligotyping pattern of SIT745 for Malaysia, remarkable by absence of spacers 37, 38 and 40 (tentatively relabeled as EAI-MYS sublineage [13]). Furthermore, an interrogation of web-tool queried as percentage by country of strains belonging to the octal code 777777777413131 (SIT745) confirmed their phylogeographical specificity for Malaysia (Supplementary File S3).

As another example of usefulness of SpolSimilaritySearch tool, one may refer to a recent study by Balcells et al., 2015 [14], focusing