Towards a better understanding of the long-lasting evolutionary history of Mycobacterium tuberculosis

Gaetan Senelle, Christophe Guyeux, Guislaine Refrégier, Christophe Sola

Tuberculosis 143 (2023) 102374
Available online 25 November 2023

Review

Towards a better understanding of the long-lasting evolutionary history of *Mycobacterium tuberculosis*

Gaetan Senelle [a], Christophe Guyeux [a,*], Guislaine Refr´egier [c], Christophe Sola [b]
[a] *Institut FEMTO-ST, UMR 6174 CNRS, Universit´e de Franche-Comt´e, France*
[b] *IAME UMR 1137 INSERM, Universit´e Paris-Saclay, France*
[c] *Laboratoire ESE, Universit´e Paris-Saclay, France*

* Corresponding author.
E-mail address: christophe.guyeux@univ-fcomte.fr (C. Guyeux).
Contents lists available at ScienceDirect

journal homepage: www.elsevier.com/locate/tube
https://doi.org/10.1016/j.tube.2023.102374
Received 21 January 2023; Received in revised form 30 June 2023; Accepted 3 July 2023

A B S T R A C T
The daily increasing sequencing of *Mycobacterium tuberculosis* has made it possible to establish an advanced phylogeny of this bacterium. It currently includes 9 lineages mainly affecting humans, completed by animal lineages, which form the *Mycobacterium tuberculosis* complex. Inherited from various historical approaches, this phylogeny is now based on Single Nucleotide Polymorphisms (SNPs), of which updates are frequently proposed. We present here evidence that the task needs refinements: some lineages have currently suboptimal defining SNPs, and many sublineages still need to be named and characterized. These findings are based on a new tool specifically designed to index the entire existing sequencing data. In this article, we focus on lineages 4.5, 4.7, 6 and 7. We take the opportunity to present some evidence that TB-annotator shows strong relevance, identifying well supported sublineages, as well as good global agreement with previous findings.

1. Introduction

Tuberculosis (TB) still remains a significant global health problem. According to the World Health Organization (WHO), TB is one of the top 10 causes of death worldwide and the leading cause of death from a single infectious agent. In 2020, an estimated 10 million people fell ill with TB, and 1.4 million died from the disease. Additionally, the emergence of drug-resistant forms of TB is a growing concern, as these strains are much more difficult to treat and can lead to higher mortality rates. The agent of TB is *Mycobacterium tuberculosis* complex (MTBC) that encompasses lineages transmitting between humans and the lineage adapted onto animals including several species, among which *M. bovis*, *M. caprae* and *M. pinnipedi*. Studying the global diversity of MTBC is important for understanding the disease, developing new treatments, and ultimately controlling and eliminating TB as a public health threat. For more than one decade, the study of this disease can also be done through the analysis of genomes obtained from numerous strains.

Having many different MTBC genomes is interesting for several reasons. Firstly, it allows researchers to build the foundations for an understanding of the genetic diversity of MTBC. MTBC encompasses several species due to its obvious ecological adaptations: some sublineages such as *M.bovis*, *M. caprae* but also *M. microti* are widely found in a single, homogeneous host taxon. *M. africanum* accounts for up to 20% of MTBC diversity in Western and Central Africa and less than 1% in Europe of the Americas despite large population migrations. By studying the entire genomes of different clinical isolates worldwide, researchers can identify the specific genetic variations that exist within the populations. This can provide insight into how MTBC evolves and spreads, and how it adapts to different host environments. Secondly, it can help researchers understand the emergence and spread of drugresistant strains of MTBC. By comparing the genomes of drug-sensitive and drug-resistant strains, researchers can identify the specific genetic changes that confer resistance to certain drugs. This can aid in the development of new drugs and treatment strategies to combat drugresistant TB, in a personalized treatment approach. Thirdly, it can help to develop new vaccine candidates. By comparing the genomes of different strains of MTBC, researchers can identify genes and proteins that are conserved across the population, and that could be targeted by a vaccine. This can help to improve the chances of developing a vaccine that is effective against a wide range of MTBC strains. Lastly, studying many different MTBC genomes can also aid in the development of more accurate diagnostic tools. A better knowledge of the genetic diversity of the bacteria will allow for a more refined identification of the disease. There are currently over 160,000 genomes available on NCBI [1].

However, many things remain to be understood about this bacterium, and being able to analyze such a database should help to enrich our knowledge. This is particularly obvious in the case of the analysis of the phylogeny of *M.tuberculosis*. If a lot of progress has been made in the knowledge of this phylogeny thanks to numerous studies comparing several hundreds or thousands of sequenced genomes, there are still many grey areas that the increase by almost two orders of magnitude of the sampling should help us to remove. Among these shadowy areas of phylogeny, we have chosen to illustrate in this article the following cases. First, lineage 6 [4] has only recently been explored by Coscolla et al. [5] but still a restricted number of samples (n = 338), not including all presently available data and not including outgroups so that several classifying SNP lack specificity. Lineage 4 [4,18], on the other hand, was heavily sequenced but is currently heterogeneous, with some sublineages highly represented and subdivided (e.g. 4.1) and others poorly characterized (4.5 and 4.7). Finally, no sublineage of Ethiopian L7 [4] is currently defined. The aim of this article is to illustrate these imperfections, and to show that with a new ad hoc tool, namely the TB-annotator, we have the means to make light into these grey areas, by evaluating the quality of currently selected Single Nucleotide Polymorphism (SNP) for lineages definition, and by searching for new clades sharing exclusive characters.

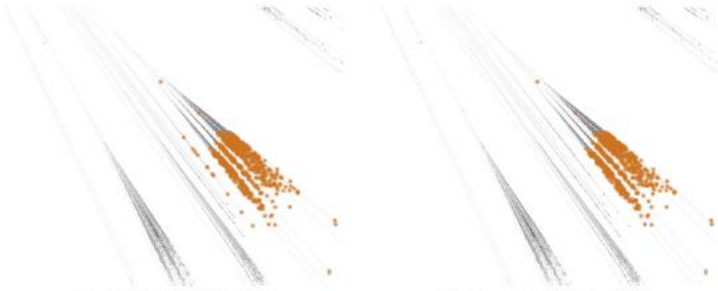
2. Material and method

The work presented here is based on our new tool called TB-annotator, in its 15901-strain version. We briefly recall here how it works. For further information, see Ref. [14]. The first step of the TB-annotator is the preparation of the reads. SeqKit [16] is first used to estimate the likely coverage of the genome under consideration. A estimated coverage of 95% and 20X depth were used as thresholds to exclude low depth and contaminated samples. Fastp [2] is then applied for quality control and the following preprocessing steps. Adapters are trimmed, and low quality bases at ends are removed. Also, reads with too many low quality bases are removed. Finally, various advanced database corrections are made. Processed reads are mapped on a reference genome using BWA-MEM [9], while duplicate reads are marked using SAMTOOLS [10], and the sequence alignment is saved in CRAM format. Note that optimization is achieved thanks to the

Snakemake [8] ability to continue partially executed workflows, and sequence alignment is used in all steps, instead of raw reads. Some statistics are finally computed with SAMTOOLS using the sequence alignment file, like mean read depth and the proportion of covered bases. This information can be used to assess the quality of the sample and the quality of sequencing. Variant calling is performed by Snippy [13]. False SNP calling close to structural variations are avoided by applying Samclip [12] on aligned reads. Variants are then called with Freebayes, and the final output is a VCF file. These variants are annotated using SnpEff [3] and identified using their SPDI notation. Regarding large insertions and deletions, a custom Python script creates the list of clipped reads, the position of the left clipping and right clipping, and the clipped sequence. Let us recall that a clipping signal is the position where a configuration amount of reads are clipped (when a read partially matches the reference sequence, the aligner separates the unmatched part). The list of clipping signals is then obtained thanks to the bedtools groupby function. Such signals allow the detection of insertion sequences (IS) and of regions of deletion (RD). The SNP lists obtained then allow to build an alignment of the considered strains, and the use of RAxML [17] (GTRCAT model) allows to build the associated phylogenetic tree. The latter is reproduced in an ad hoc web interface, in which it is possible to select a set of strains and to visualize the characters (SNPs, IS, RD …) that they have in common, and exclusively. An exclusive character of a clade is a good candidate to be a phylogenetic marker. Conversely, a SNP known to be a phylogenetic marker of a given sublineage, but which does not appear to be exclusive (either because it is present elsewhere in the tree, or because strains in the selection do not have it), is not a good marker. Individual quality of each SNPs can be visualized on the tool, as well as available metadata of the samples. TB-annotator [14] is now online,[1] but with private access. Readers interested in accessing this analysis platform can send a request for access to the authors of this article.

3. Results

Analysis of the 15,901 strains by our pipeline led to the detection of 180 RDs and more than 300,000 good quality SNPs. This list contained the SNPs proposed by Coscolla [4] to define the L6 sublineages, those of Stucki [18] for the L4 sublineages, and the SNPs of Coll [4]. This allowed us to question the quality of these SNPs based on our 15,901 strains. We also took the opportunity to interrogate two more recent lineage characterization proposals, those of Napier [11] and Freschi [6]. Regarding L6, we first recall that it was defined by SNP C1816587G in Coll et al. (2014), but that the team did not propose any sublineage. Coscolla and coworkers, for their part, identified three sublineages, numbered 6.1, 6.2, 6.3, each having in turn 3 sublineages. While they did propose SNPs for sublineages 6.1.1 to 6.3.3, they did not propose anything for L6, nor for 6.1 to 6.3. Napier et al., on the other hand, proposed 10 new SNPs for defining L6, but did not retain Coll's. They also did not define any sublineages. Freschi, finally, returned to the Coll definition of L6 (C1816587G), and with no sublineages. The Coll et al. SNP of L6, first, was identified in 609 strains in TB-annotator tool, and this selection showed two exclusive variants. However, looking in detail, the strains selected by this SNP are L6 as well as L9, see Fig. 1a. On our side, we found that SNP G41241A allows to define exclusively L6 (excluding L9): it is present in 599 strains that share 78 exclusive variants (Fig. 1b). Exploring deeper Coscolla's proposal, 6.3.1 seems to be well defined, but not 6.3.2 and 6.3.3 do not: the number of strains presenting the different SNP presented as specific varies from simple to double depending on the SNP chosen (several SNPs are proposed for the same sub-lineage). Regarding Coscolla's 6.2, this sublineage was well supported by TB-annotator but the SNP of 6.2.2 also turned on 6.2.1 and 6.2.3. Similar observations were made for 6.1: 6.1.1 and 6.1.3 SNPs are well supported by TB-annotator, but 6.1.2 was poorly defined, including 2 strains 6.1.3. Finally, ten isolates from 6.1's belong to no sublineages, although they seem to form clades. Concerning L7, Freschi's and Napier's definitions match Coll's definition of G1137518A. However, this SNP achieves only 97% exclusivity. In contrast, TB-Annotator revealed 836 variants that are 100% exclusive. This is the case for G1960C, which we propose as the best feature of L7. Moreover, there is currently no defined sublineage of L7, whereas with our interface we gathered evidence that at least 3 sublineages could be defined. The first one consists of at least 10 strains, and has various exclusive variants among which T109162C. Similarly, G160716A defines a group of 18 strains and G1932863T a group of at least two strains. More advanced studies regarding the quality of these SNPs and associated clades are needed to confirm their relevance. Concerning lineage 4, as we have seen, some sublineages are much more thoroughly characterized than others. 4.7, for example, has no sublineage in the various phylogenies currently proposed, and Stucki, Freschi, and Napier all use Coll's definition (C4249732A). Using the TB-annotator, we first found a better characterization of 4.7 (C10741G), and many sublineages appear to be definable from SNPs. A potential scheme is reproduced in Fig. 2, which will require confirmation in future work. Exactly the same is true for 4.5, where the list of sublineage SNPs could be the one in Table 1.



(a) Coll's C1816587G (b) Our G41241A SNP

Fig. 1. Coll's SNP of L6 encompasses both L6 and L9. As can be seen in the figure on the left, the SNP proposed by Coll to define L6 also contains strains below the clade grouping all L6s. On further investigation, these lower strains turn out to be L9. Our SNP separates L6 from L9, as shown in the figure on the right.

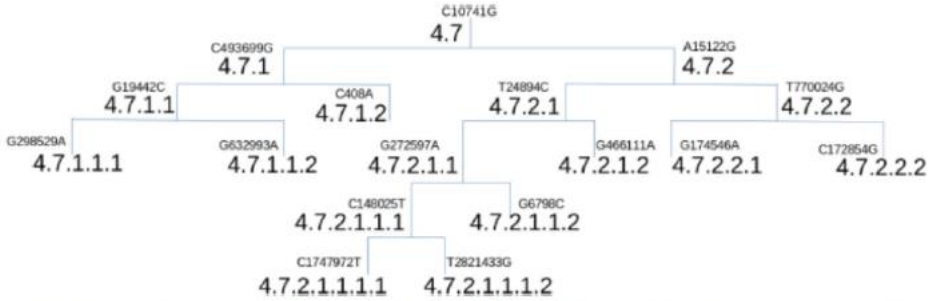


Fig. 2. A proposed definition of sub-lineages for L4.7. This original subdivision of L4.7 into sublineages was obtained in TB-annotator by searching for significant clades with at least one exclusive character.

Table 1
Possible SNPs for a 4.5 phylogeny. These SNPs were found using TB-annotator: for each significant clade in the tree, we looked to see if there was at least one exclusive character. If so, we proposed a definition of sublineage, with the exclusive SNP as a characteristic.

Lineage	SNP
451	275523
452	275624
453	212629
454	570543
4511	85961
4512	501530
4513	3879070
45131	39725
45132	255444
4531	316970
4532	128589
45321	4294863
45322	1853597
4541	194738
4542	4314
45421	384635
45422	385778
4543	2331267

4. Discussion

The most detailed characterization of lineage 6 is for the moment the one proposed by Coscolla. The 3 main sub-lineages they define seem to be true according to our study, as well as the existence of sublineages for 6.1 to 6.3. However, it is regrettable that there are only features for the third level of classification, and that neither the L6 nor the 6.1 to 6.3 have a defining SNP or SNP list. With our interface, many candidates (exclusive variants) appear, like C3887907A for the 6.1, G4393A for the 6.2, and C98536T for the 6.3. These candidates will need to be investigated further to gain a better understanding of L6. Similarly, it could be seen that better SNPs can be defined for each of the 6.1.1 to 6.3.3 sublineages, reducing or negating false positives (strains outside the clade that still have the SNP) and false negatives (strains in the clade without the SNP). Finally, it appears to us that these sublineages can in turn be divided into sublineages, which will be the subject of further study in the future. L7, on the one hand, is clearly understudied. We have seen that the SNP that has defined it for almost ten years is not optimal. In addition, it seems quite possible to define sub-lineages in L7, but this would probably require a larger number of genomes to be confirmed. Last, we showed that lineages 4.5 and 4.7 are both currently poorly characterized (presence of false positives and false negatives), and poorly decomposed (no sub-lineage definitions), although such definitions are quite possible. The objective of this work was not to propose a new phylogeny, but to raise many points showing, on the one hand, that the current characterizations are perfectible, and on the other hand, that we can probably go further in the definition of certain sublineages with new tools using the entire sequencing data currently available. These improvements, achieved for instance in Refs. [7,15], are made possible by the very large number of genomes that we now have at our disposal, but they require the development of ad hoc tools to manage such large amounts of data, and to produce the analyses necessary for these new definitions. TB-annotator is such a tool, which we are intensively developing. A first version is currently available on demand. All computations have been performed on the Mesocentre de Franche-Comt´e supercomputer facilities.

Contribution Conceptualization: CS, GR, CG; Data curation: GS, CS; Formal analysis: GS, GR; Funding acquisition: CG; Investigation: CS, CG; Methodology: GR, GS, CS, CG; Project administration: CS; Resources: CG; Software: GS; Supervision: CS; Validation: GR; Visualization: GS; Roles/Writing – original draft: CG; Writing – review & editing: CG, CS.

Declaration of competing interest None.

Transparency declaration This article is part of a supplement entitled *"Paleopathology and Evolution of Tuberculosis" - Conference Proceedings from the 3rd International Congress on the Evolution and Paleoepidemiology of Tuberculosis (ICEPT-3)* published with support from the K 125561 ('Tuberculosis and Evolution') research grant of the National Research, Development and Innovation Office (NKFIH - Hungary) and the Department of Biological Anthropology, University of Szeged, Hungary.

References [1] [Internet] National center for biotechnology information (NCBI). Bethesda (MD): National library of medicine (us), national center for biotechnology information; 1988 [cited 2022 nov 18]. Available from: https://www.ncbi.nlm.nih.gov/. [2] Chen Shifu, Zhou Yanqing, Chen Yaru, Gu Jia. fastp: an ultra-fast all-in-one fastq preprocessor. Bioinformatics 2018;34(17):i884–90. [3] Cingolani Pablo, Platts Adrian, Wang Le Lily, Coon Melissa, Nguyen Tung, Wang Luan, Land Susan J, Lu Xiangyi, Ruden Douglas M. A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. Fly 2012;6(2): 80–92. [4] Coll Francesc, McNerney Ruth, Guerra-Assunç˜ao Jos´e Afonso, Glynn Judith R, Perdig˜ao Jo˜ao, Viveiros Miguel. Isabel Portugal, Arnab Pain, Nigel Martin, and Taane G Clark. A robust snp barcode for typing mycobacterium tuberculosis complex strains. Nat Commun 2014;5(1):1–5. [5] Coscolla Mireia, Gagneux Sebastien, Menardo Fabrizio, Loiseau Chlo´e, Ruiz- Rodriguez Paula, Borrell Sonia, Darko Otchere Isaac, Asante-Poku Adwoa, Asare Prince, S´anchez-Bus´o Leonor, Gehre Florian, N'Dira Sanoussi C, Martin Antonio, Affolabi Dissou, Fyfe Janet, Beckert Patrick, Niemann Stefan, Alabi Abraham S, Grobusch Martin P, Kobbe Robin, Parkhill Julian, Beisel Christian, Fenner Lukas, B¨ottger Erik C, Meehan Conor J, Harris Simon R, Bouke C, de Jong, Yeboah-Manu Dorothy, Brites Daniela. Phylogenomics of mycobacterium africanum reveals a new lineage and a complex evolutionary history. Microb Genom 2021;7(2). [6] Freschi Luca, Vargas Roger, Husain Ashaque, Kamal SM, Alena Skrahina, Tahseen Sabira, Ismail Nazir, Anna Barbova, Niemann Stefan, Cirillo Daniela Maria, et al. Population structure, biogeography and transmissibility of mycobacterium tuberculosis. Nat Commun 2021;12(1):1–11. [7] Guyeux Christophe, Senelle Gaetan, Refr´egier Guislaine, Bretelle-Establet Florence, Cambau Emmanuelle, Sola Christophe. Connection between two historical tuberculosis outbreak sites in Japan, honshu, by a new ancestral mycobacterium tuberculosis l2 sublineage. Epidemiol Infect 2022;150. [8] K¨oster Johannes, Rahmann Sven. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics 2012;28(19):2520–2. [9] Li Heng. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. 2013. *arXiv preprint arXiv:1303.3997*. [10] Li Heng, Handsaker Bob, Wysoker Alec, Tim Fennell, Ruan Jue, Homer Nils, Marth Gabor, Abecasis Goncalo, Durbin Richard. 1000 Genome Project Data Processing Subgroup, et al. The sequence alignment/map (sam) format and samtools. Bioinformatics 2009;25(16):2078–9. [11] Gary Napier, Campino Susana, Merid Yared, Abebe Markos, Woldeamanuel Yimtubezinash, Abraham Aseffa, Hibberd Martin L, Phelan Jody, Clark Taane G. Robust barcoding and identification of mycobacterium tuberculosis lineages for epidemiological and clinical studies. Genome Med 2020;12(1):1–10. [12] Seemann T. Samclip: filter sam file for soft and hard clipped alignments. 2020. [13] Torsten Seemann. Snippy: fast bacterial variant calling from ngs reads. 2015. [14] Senelle Gaetan, Guyeux Christophe, Refr´egier Guislaine, Sola Christophe. Tb-annotator: a scalable web application that allows in-depth analysis of very large sets of publicly available mycobacterium tuberculosis complex genomes. bioRxiv; 2023. [15] Senelle Gaetan, Rabiu Sahal Muhammed, Kevin La, Molina-Moya Barbara, Dominguez Jose, Panda Tukur, Cambau Emmanuelle, Refregier Guislaine, Sola Christophe, Guyeux Christophe. An updated evolutionary history and taxonomy of mycobacterium tuberculosis lineage 5, also called m. africanum. bioRxiv; 2022. [16] Shen Wei, Le Shuai, Li Yan, Hu Fuquan. Seqkit: a cross-platform and ultrafast toolkit for fasta/q file manipulation. PLoS One 2016;11(10):e0163962. [17] Stamatakis Alexandros. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 2014;30(9):1312–3. [18] Stucki David, Brites Daniela, Jeljeli Leïla, Coscolla Mireia, Liu Qingyun, Trauner Andrej, Fenner Lukas, Rutaihwa Liliana, Borrell Sonia, Luo Tao, et al. Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages. Nat Genet 2016;48(12):1535–43.