Towards the reconstruction of a global TB history using a new pipeline “TB-Annotator”

Gaetan Senelle, Muhammed Rabiu Sahal, Kevin La, Typhaine Billard-Pomares, Julie Marin, Faiza Mougari, Antoine Bridier-Nahmias, Etienne Carbonnelle, Emmanuelle Cambau, Guislaine Refrégier, et al.

Tuberculosis 143 (2023) 102376

Contents lists available at ScienceDirect

Tuberculosis

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

Review

Towards the reconstruction of a global TB history using a new pipeline "TB-Annotator"

Gaetan Senelle [a,1], Muhammed Rabiu Sahal [b,c,1], Kevin La [c,e], Typhaine Billard-Pomares [d,g], Julie Marin [d,g], Faiza Mougari [c,e], Antoine Bridier-Nahmias [c], Etienne Carbonnelle [c,d,g], Emmanuelle Cambau [c,e], Guislaine Refrégier [b,f], Christophe Guyeux [a], Christophe Sola [b,c,*]

[a] FEMTO-ST Institute, UMR 6174, CNRS-Université Bourgogne Franche-Comté (UBFC), France
[b] Université Paris-Saclay, 91190, Gif-sur-Yvette, France
[c] Université Paris-Cité, IAME, UMR 1137, INSERM, Paris, France
[d] Service de microbiologie clinique, Hôpital Avicenne, 93017, Bobigny, France
[e] AP-HP, GHU Nord site Bichat, Service de mycobactériologie spécialisée et de référence, Paris, France
[f] Ecologie Systématique Evolution, Université Paris-Saclay, CNRS, AgroParisTech, UMR ESE, 91405, Orsay, France
[g] Université Paris 13, IAME, Inserm, 93017, Bobigny, France

A B S T R A C T

Mycobacterium tuberculosis complex (MTBC) has a population structure consisting of 9 human and animal lineages. The genomic diversity within these lineages is a pathogenesis factor that affects virulence, transmissibility, host response, and antibiotic resistance. Hence it is important to develop improved information systems for tracking and understanding the spreading and evolution of genomes. We present results obtained thanks to a new informatics platform for computational biology of MTBC, that uses a convenience sample from public/private SRAs, designated as *TB-Annotator*. Version 1 was a first interactive graphic-based web tool based on 15,901 representative genomes. Version 2, still interactive, is a more sophisticated database, developed using the Snakemake Workflow Management System (WMS) that allows an unsupervised global and scalable analysis of the content of the USA National Center for Biotechnology Information Short Read Archives database. This platform analyzes nucleotide variants, the presence/absence of genes, known regions of difference and detect new deletions, the insertion sites of mobile genetic elements, and allows phylogenetic trees to be built, imported in a graphical interface and interactively analyzed between the data and the tree. The objective of *TB-Annotator* is triple: detect recent epidemiological links, reconstruct distant phylogeographical histories as well as perform more complex phenotypic/genotypic Genome-Wide Association Studies (GWAS). In this paper, we compare the various taxonomic SNPs-based labels and hierarchies previously described in recent reference papers for L1, and present a comparative analysis that allows identification of *alias* and thus provides the basis of a future unifying naming scheme for L1 sublineages. We present a global phylogenetic tree built with RAxML-NG, and one on L2; at the time of writing, we characterized about 200 sublineages, with many new ones; a detail tree for Modern L2 and a hierarchical scheme allowing to facilitate L2 lineage assignment are also presented.

1. Introduction

Mycobacterium tuberculosis complex (MTBC) designates a mammal and human bacterial pathogen that is responsible of tuberculosis (TB), one of the greatest scourge of human kind, that remains a model of social infectious disease, i.e. a disease that, as Covid19, points to the injustice and inability of human kind to collectively achieve an ideal of minimal wealth for everybody on earth. Whether TB was initially an animal or a human disease remains debated. Epidemiologists believed that animals initially have infected humans, whereas scientists working on genomes in 2001 suggested a different scenario: MTBC could have been hosted primarily by anatomically modern humans. At least nine human lineages and many other animal lineages have been described so far (for a review see [1]). During the last decade, the methods that are able to characterize whole bacterial genomes at high-throughput, the so-called *next-Generation-Sequencing* (NGS) methods, have allowed an improved spatio-temporal tracking of tuberculosis transmission. Among additional results, a better knowledge on the global historical and spatial spreading history of TB have been proposed. However, the comparison of hundreds of thousands of genomes remains a difficult task. Among the difficulties is the screening of individual-specific versus lineage-specific variations, as well as the understanding of the consequences of these variations on the physiology of the bacillus and on its virulence, and on the evolution of virulence across the times. It is also of particular and renewed interest

* Corresponding author. Université Paris-Saclay, 91190, Gif-sur-Yvette, France.
E-mail address: christophe.sola@universite-paris-saclay.fr (C. Sola).
[1] These two authors contributed equally

https://doi.org/10.1016/j.tube.2023.102376
Received 18 January 2023; Received in revised form 3 July 2023; Accepted 10 July 2023
Available online 25 November 2023

Table 1
A benchmark analysis of sublineage names of L1, in the classification of Coll (C), Napier (N) Freschi (F), Palittapongarnpim (P), with indicators of representativity increase according to databases growth evolution. From left to right: sublineage names, aliases, marker positions, total number of genomes in the v.1 database, number of exclusive markers and of markers 95% exclusive, total number of the same sublineage in the v.2.1 database, increase factor in genome number of a sublineage (v2.1/v1), relative increase between v1 and v2, v2/v1 growth rate, "sequencing effort" as defined by the relative increase between v1 and v2 divided by the increase factor. Green cells show sublineage growing slowlyer than increase in sequencing, red cells, are those growing faster.

Lineage	Alias	Position	Ref->var	Genome n°, Exclusive, 95% (version 16000)	v2.1* (90500 SRAs)	Increase factor (v2/V1)	increase in DB	sequencing effort
L1-CN		615938	G->A	(1560, 1, 468)	6524	4.18	5.625	0.74
L1-F		64028	C->T	(1577, 1, 467)	6686	4.24	5.625	0.75
L1.1-CN		4404247	G->A	(1037, 1, 31)	3551	3.42	5.625	0.61
L1.1-F		2989683	C->T	(1039, 1, 31)	3546	3.41	5.625	0.61
L1.1.1-C		3021283		(745, 1, 103)	1343	1.80	5.625	0.32
L1.1.1-N	L1.1.1-F	529363	C->T	(744, 21, 89)	1338	1.80	5.625	0.32
L1.1.1.1-C		3216553		(373, 41, 33)	698	1.87	5.625	0.33
L1.1.1.1-N		1750465	T->C	(373, 41, 33)	690	1.85	5.625	0.33
L1.1.1.1-F		784450	C->T	(672, 4, 43)	1198	1.78	5.625	0.32
L1.1.2-CN		2622402	G->A	(85, 32,77)	1543	18.15	5.625	3.23
L1.1.2-F		1831531	G->A	(86, 28, 82)	1534	17.84	5.625	3.17
L1.1.4-G		22815		(25, 178, 0)	33(SPDI:25)	1.00	5.625	0.18
L1.1.3-C		1491275		(255, 1, 54)	708	2.78	5.625	0.49
L1.1.3.1-N		403481	C->T	(99, 1, 89)	335	3.38	5.625	0.60
L1.1.3.2-N	L1.1.1.2.i1-F	285096	G->T	(101, 3, 298)	200	1.98	5.625	0.35
L1.1.3.i1-F		3389975	T->C	(100, 40, 256)	173	1.73	5.625	0.31
L1.1.3.3-N		2738352	C->T	(25, 100, 42)	91	3.64	5.625	0.65
L1.1.1.2-F/L1.1.3-CN	L1.1.3-CN	15177	C->G	(255, 0, 55)	710	2.78	5.625	0.49
L1.1.1.2-F		1167274	C->T	(23, 95, 20)	61	2.65	5.625	0.47
L1.2.1-C		3479545		(414, 0, 397)	1869	4.51	5.625	0.80
L1.2.1-N		590595	G->A	(9, 104, 0)	41	4.56	5.625	0.81
L1.2.1-F		1136017	A->G	(415, 60, 252)	1871	4.51	5.625	0.80
L1.2.1.1-F/L1.2.2-N	L1.2.2-N	16526	C->T	(407, 1, 314)	1791	4.40	5.625	0.78
L1.2.1.1.1-F		2737201	A->C	(358, 33, 23)	1525	4.26	5.625	0.76
L1.2.1.1.2-F		405404	C->T	(48, 1, 28)	259	5.40	5.625	0.96
L.1.2.1.2-F		4237703	A->G	(10, 5, 0)	77	7.70	5.625	1.37
L1.2.2-C		3470377		(161, 5, 81)	1136	7.06	5.625	1.25
L1.2.2-N		528781	G->A	(405, 155, 162)	1787	4.41	5.625	0.78
L1.2.2-F/1.2.2-C/L1.3-N		2763624	G->A	(162, 2, 82)	1135	7.01	5.625	1.25
L1.2.2.1-F/L1.3.2-N		61842	T->C	(130, 15, 97)	1002	7.71	5.625	1.37
L1.2.2.2-F/L1.3.1-N		1237815	G->A	(32, 1, 52)	140	4.38	5.625	0.78
L1.3.1-N		1245275	C->T	(32, 1, 52)	140	4.38	5.625	0.78

C= Coll et al. 2014; N=Napier et al. 2020 ; F=Freschi et al. 2021, G-Guyeux, *as assessed on Nov. 23rd

to understand the long-term co-evolution of the pathogen and its hosts following the ecological, demographical, and migratory changes.

Even if bioinformatics pipelines have already been developed by many teams of bioinformaticians around the world, we undertook the "*TB-Annotator*" project, whose aim is to develop a multiscale system, *i.e.* an information system able to detect recent TB outbreaks, as well as to decipher the long-term TB evolution; one of our ambition was to develop a new pipeline that would integrate the previous genome knowledge acquired during decades on repetitive-DNA based system and on regions of differences, to the new SNPs-based and gene content schemes, thus enhancing the resolutive power of the analysis pipelines.

2. Material and methods

2.1. Creation and implementation of TB-annotator

We present results obtained with two versions of *TB-Annotator*. *TB-Annotator* v.1 was developed between 2020 and 2021 by creating an initial database of 15901 Sequence Read Archives (SRAs), that was compiled from two sources. On the one hand, NCBI *M. tuberculosis* specific bioprojects were searched [2], and those that allowed a good geographical representativeness were retained. On the other hand, a constant number of SRAs were randomly extracted from each of Freschi *et al.* sublineages [3] which represents for the moment, together with Napier *et al.,* one of the most comprehensive acknowledged and published list of lineages and sublineages [4]. These SRAs were then indexed according to their mutations with respect to the H37Rv reference sequence (NC_000962.3). During this work, we also integrated the recent new L2 sublineages nomenclature that was published by Thawornwattana *et al.* [5].

TB-Annotator v.2 is an unsupervised system that collects all SRAs from the NCBI and assess their quality. On July 3rd, we are currently working on v2.3 that contains 102,001 SRAs.

After being downloaded, the SRAs are processed through *fastp* processing [6], the process including quality profiling, filtering of bad reads, trimming, cutting of low quality bases, and cutting of adapters. Mapping is then performed by BWA [7] and SNPs and Indels search using *snippy* [8]. Detection of difference regions and Insertion Sequences is based on clipping signals and coverage elements, obtained using new Python scripts. Spoligotype is reconstructed based on a *blast* process against 68 spacers and scoring. Drug resistance type are obtained using TB-Profiler [9]. Phylogenetical trees are built using RAxML-NG (global tree on 15,901 SRAs, v1, sublineage-specific trees in the v2.1 version) [10,11]. The pipeline is more exhaustively described in Ref. [12].

3. Results

3.1. Benchmark of most recent global SNPs-based classification schemes for L1 and attempt to unify designations

We compared results on L1 SNP-based markers obtained in various studies and provide a comparison between results obtained by investigators. Table 1 provides the results of this comparison, that allows to provide translation between *alias* names given to identical sublineages by different investigators, as published in some landmarks papers [3,4,13].

From Table 1, we notice that investigators chose different SNPs to define the same lineage (even if they never discuss the *informativeness* of the chosen markers and in some cases some could be suboptimal), whereas in other cases, different investigators named identical

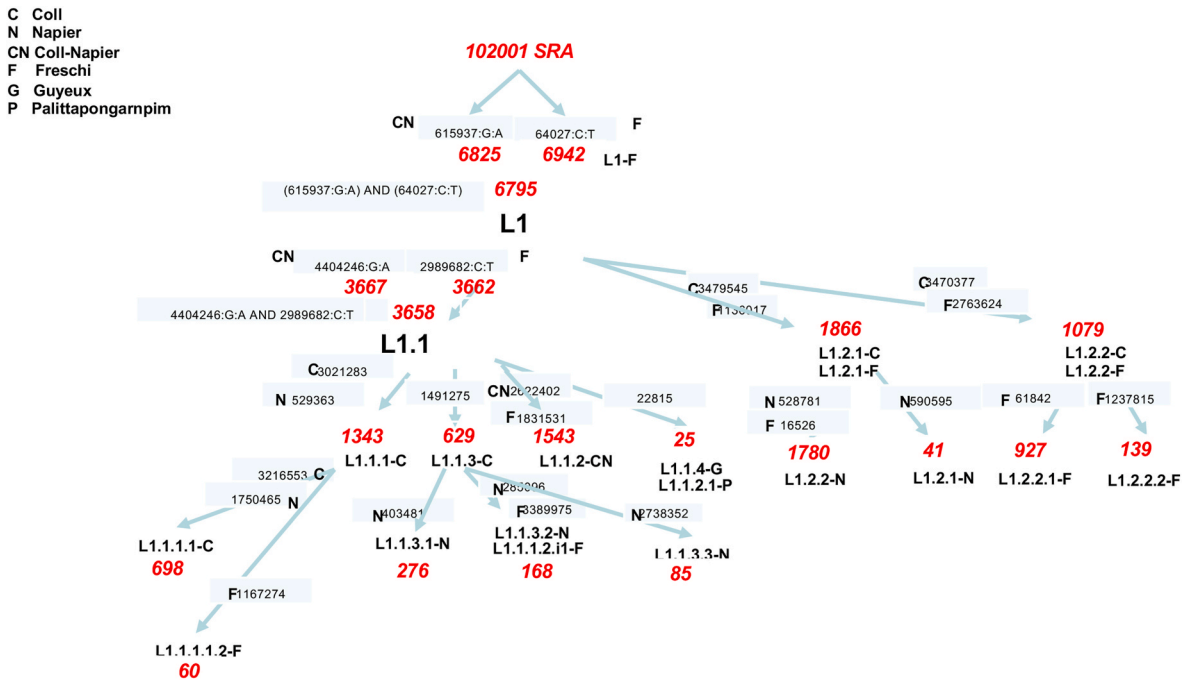

Fig. 1. A hierarchical SNP-based algorithm to subclassify L1 isolates based on four previous landmark papers (Coll *et al.*, 2014, Palittapongarnpim *et al.*, 2018, Napier *et al.*, 2020, Freschi *et al.*, 2021). Total SRA n° (italicized red characters) according to TB-Annotator v2.3, 102,001 genomes (depending on SNP combinations used to screen SRAs, and on undesignated branches, the total number of SRA in branches may not always coincide with the upper hierachical branch if some isolates remain unclassified. For each lineage designation and SNP, a letter (C,N,F,P,G) follows or preceeds the naming to identify the investigator having suggested this marker as diagnostic. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

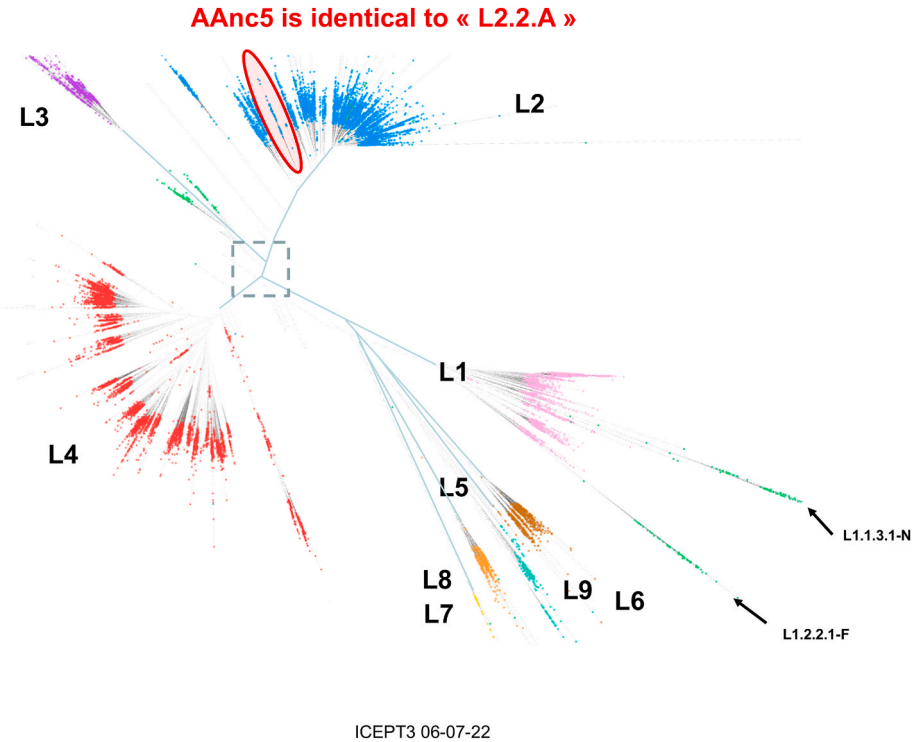

Fig. 2. RAxML-NG built phylogenetic tree on 15,901 SRAs from public and private genome collection; recent L1 subbranches described in Bangladesh, L1.1.3.1-N and L1.2.2.1-F appear in green and in a peculiar position that will have to be investigated into more details in future *TB-Annotator* versions. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

sublineages with different names (*alias)* and using different hierarchical levels; such naming creates a great deal of confusion, which could be detrimental to the discovery of recent or historical transmission cases. All *alias* names are provided, as well as an index, named "sequencing effort" that shows the relative increase in SRAs for Lineage 1 when comparing the increase of TB-Annotator v.1 versus v.2. It appears from this index, that L1.1.2 and L1.2.2 are the two sublineages that have seen a large improvement in their description between *TB-Annotator v.1* and

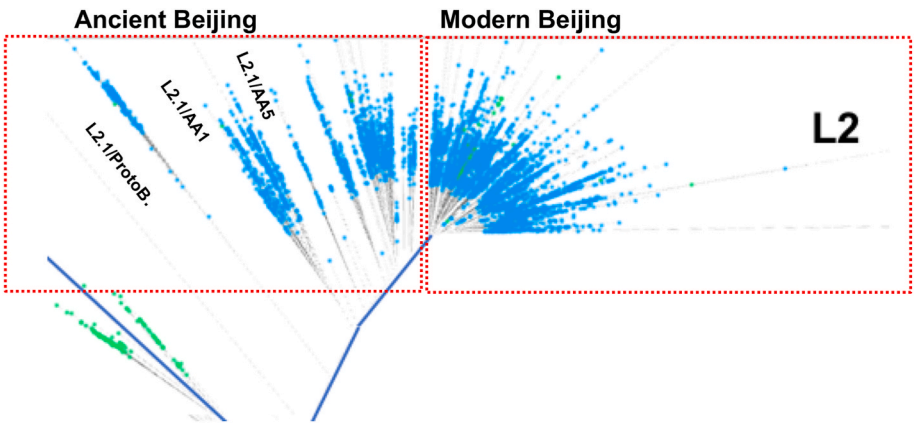

Fig. 3. Close-up on the L2 lineage, the distinction between Ancient and Modern Beijing is shown as well as some historical ancient Beijing sublineages: L2.1 (protobeijing), Asian Ancestral 1 (L2.2.AA1), Asian Ancestral 5 (L2.2.A, Japan) according to the new nomenclature of Thawornwattana et al., 2021.

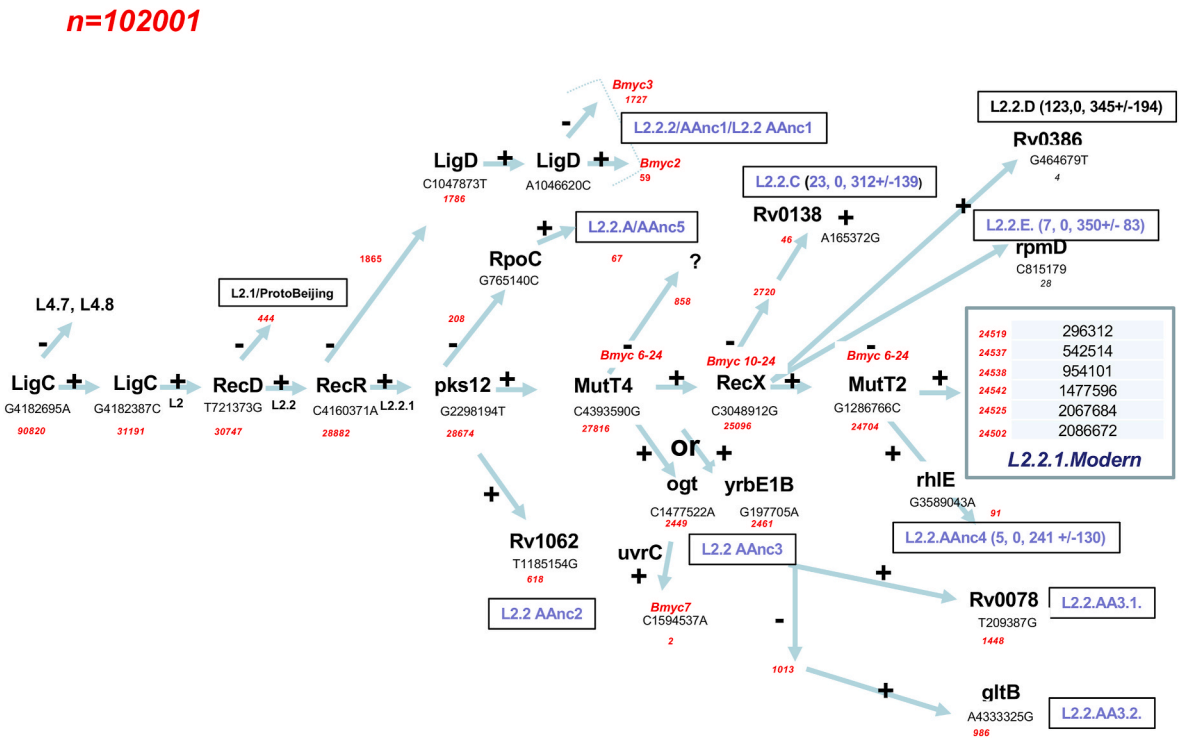

Fig. 4. Algorithm used to classify Ancient Beijing isolates based mainly of diversity of 3R genes, SNPs variants and position (H37Rv: NC_000962.3 as a reference) are found just below the genes. Lineage names in boxes are given according to the latest classification (Thawornwattana *et al.*, 2021), other names (Bmyc) are sometimes indicated and refer to Mestre *et al.*, 2011, total SRA found for a sublineage are mentioned in red or grey for L2.2.D and L2.2.E (TB-Annotator v2.3; 102,001 SRA, june 2023); the far right (L2.2.1) provides a couple of markers to diffentiate inside modern L2. In brackets: number of 100% exclusive SNPs, number of 95% exclusive SNPs, average SNP distance between isolates within the sublineage.

TB-Annotator v.2. Based on this table, a new hierarchical classification scheme was built (Fig. 1, updated on 102,001 SRAs on July 3rd, 2023). We suggest that L1.1.2.1, as named by Palittapongarnpim *et al.* should better be identified as a potential new L1.1.4, since it does not harbor, although belonging to the L1.1 sublineage, any of the SNPs characteristics of L1.1.1, L1.1.2, or L1.1.3.

3.2. Building a new global phylogenetic tree and digging into knew lineages

We built a global phylogenetic tree produced on a comparison of 15901 SRAs, that is shown in Fig. 2.

In this figure, the classical color-coded scheme for the main lineages is used (L1: pink, L2: blue, L3: purple, L4: red, L5 is in dark orange, L6: brown, L7-L8: yellow and turquoise, L9 and animal: turquoise). A specific red frame with pink color shows the recently described "Asia Ancestral 5" L2 sublineage [14], shown to be identical to L2.2A according to Ref. [5]. Two long branches within L1 (light green, L1.2.2.1-according to Freschi *et al.* for the inferior one and L1.1.3.1 according to Napier *et al.* for the superior one) and 2 shorter branches within L3 (light green too) are specific isolates from Bangladesh; their position remains unexplained and it was impossible to find specific SNP markers for these branches. We suggest that these two L1 light green branches are either linked to contemporary epidemic lineages, that are fast evolving as compared to the other ones, hence the long branches, or their position in the tree could be misleading; more analysis should be made on these branches. Going further in the analysis, we will now focus, as an example, on a analysis of L2 lineage diversity (Fig. 3).

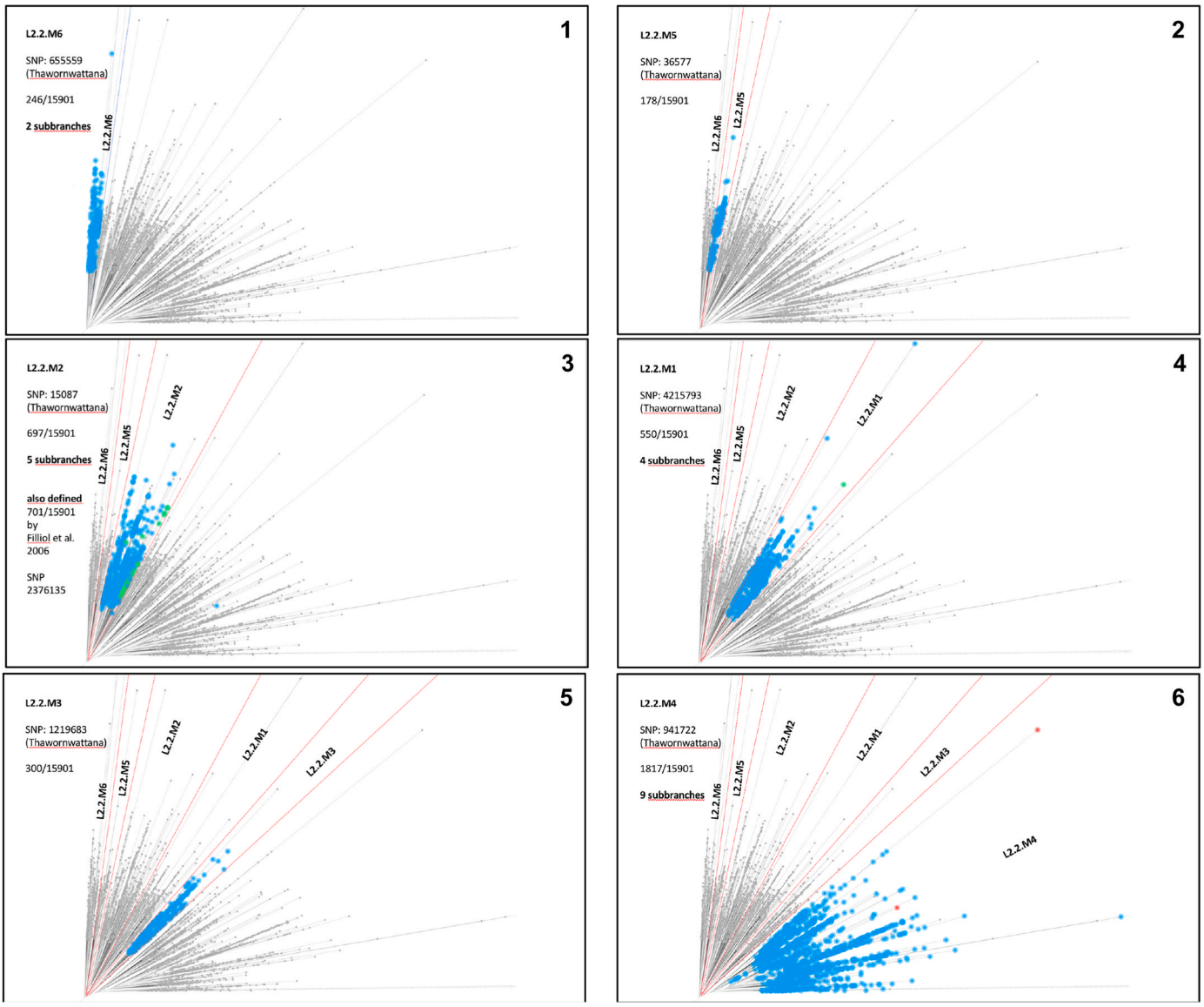

Fig. 5. Analysis of the main Modern L2 Beijing branches observed in TB-Annotator v.1 (15900 SRAs); in each frame (n = 6) are given: the name of the modern L2 sublineage, the variant position relative to H37Rv (NC_000962.3) to define this sublineage, the author having decribed this sublineage for the first-time, and the absolute number of SRAs belonging to this lineage within the TB-Annotator v.1 database (15901 SRA), and lastly, the number of subbranches defined in this sublinage.

Fig. 3 shows the specific branch of Lineage 2/Beijing, one of the most widespread lineage that has also more risks to be found as multi-drug resistant. For a long time, this lineage was shown to be difficult to study, however since the unifying study of Shitikov et al. the structure of this lineage is now understood [15]. Even if some sublineages still keep their name of origin (W148, Central Asian outbreaks, Asia African …), a new nomenclature was suggested proposed by a Thai group [5]. Two parts are clearly seen in Fig. 3. On the left of Fig. 3 are shown: (1) the ancestral Beijing (red frame of the left), with, from left to right: the L2.1 or Proto-Beijing lineage, the Asia-Ancestral 1 and the Asia Ancestral 5 from Japan (now designated as AA2) and other remaining ancestral sublineages that will be described elsewhere, (2) the modern Beijing (right red frame). The way to obtain a thorough classification of the L2 ancestral isolates may also be achieved using the new algorithm that we suggest in Fig. 4 (updated on *TB-Annotator* v.2.3 on July 3rd with 102,001 SRAs).

Fig. 4 presents a progressively hierarchical framework to obtain a thorough classification of L2 ancestral isolates, based mainly (but not only) on variants observed on 3R genes, either present (+) or absent (-). Mutations are written in small characters under the mentioned genes with the allelic change (e.g. for MutT4: a C-> G at position 4393590). In red colors also are mentioned the ancient *Bmyc* namings according to Mestre *et al.* [16]. Presence (+) or absence (-) of the different mutations allows a progression into a fine assignment according to a PHRANA scheme (Progressive hierarchical resolving assay using nucleic acids) [17]. Numbers may differ when the SNP combination required to achieve the level of classification provides a different result than the SNP search done alone. In blue boxes are given the names of the ancient Beijing as mentioned in Ref. [5]; mean intra-cluster distance in SNP is also provided for some sublineages. Other minor subbranches remain to be discovered. On the extreme right of Fig. 4 is mentioned the L2.2.1 modern clade, that requires a specific focus and specific classification.

Fig. 5 presents the 6 Modern L2 branches as they appear in the phylogenetic tree. On the top left (1), is L2.2 M6, made up of two branches (246 genomes in TB-Annotator v.1); in (2) is the L2.2 M5, followed in (3) by L2.2 M2 (5 branches), by L2.2 M1 in (4) (4 branches), by L2.2 M3 in (5) and last by L2.2 M4 (9 branches).

The Lineage 3 genetic diversity was described in detail in two recent articles [18,19] and we won't go into detail of the history and taxonomy

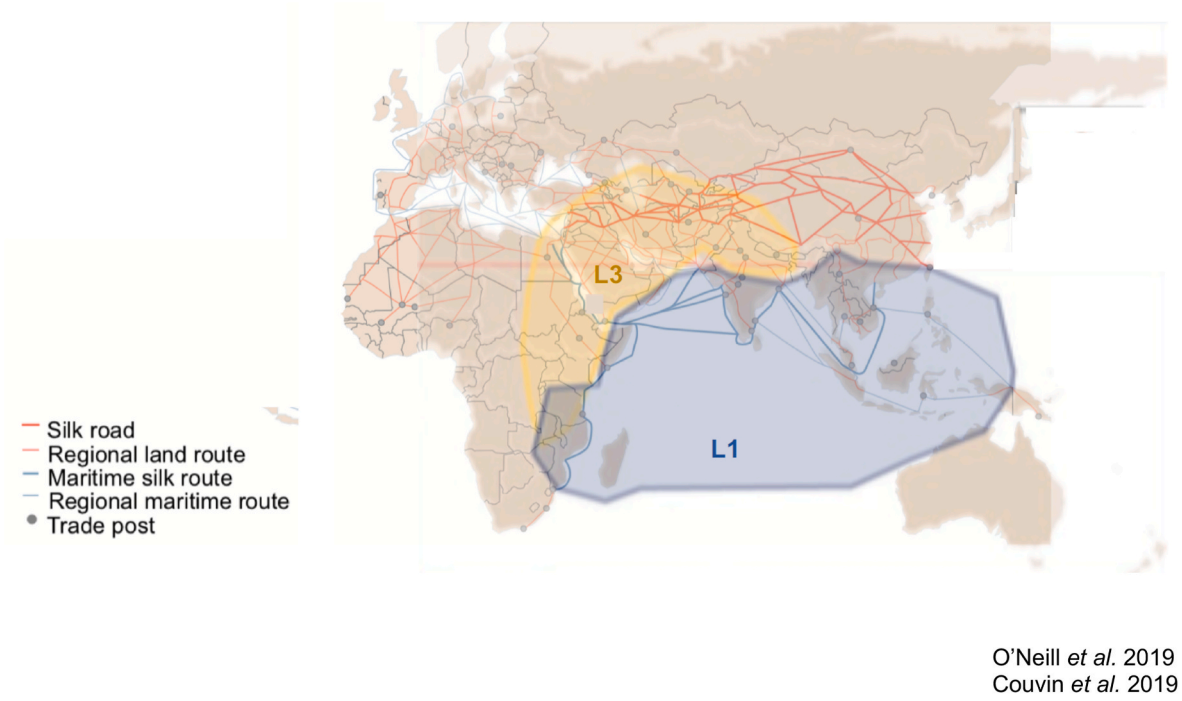

Fig. 6. Superimposition of L1 and L3 global pattern of distribution based on spoligotyping with known historical trade routes [34,35].

of this lineage. The future version of *TB-Annotator* should allow the reconstruction of a more accurate evolutionary history of all MTBC sub-lineages including Lineage 4, and to define new sublineages. A specific study describing new results obtained on the taxonomy of the Lineage 5 is also to be published elsewhere.

4. Discussion

Reconstructing the history of the worldwide spread of MTBC remains a challenge. Co-evolution between host and pathogen may have lasted between 7 and 70,000 years for tuberculosis [20–22]. We know that the diversity and severity of the clinical syndrome of tuberculosis depend on the lineage affecting the patient [23–25]. Menardo *et al.* have recently suggested that the *esx* gene, an important virulence factor related to Type VII Secretion Systems, has been under positive selection in L1[18]. This study showed that 17 genes were under positive selection, of which five were in common between L1 and L3 [18]. For L2 infections, specific SNPs are known to be under selective pressure and specific virulence factors have been discovered [26]. With the increasing amount of public WGS, we expect to delve into an "*archeological genomics*" of TB, and use *TB-Annotator* to study the link between genotypes and phenotypes and look for selection pressures on important drug-resistance or other metabolically important genes. Among these, polyketide synthase enzymes (*pks*), the enzymes necessary to build the complex cell wall of MTBC, composed of mycolic acids, phthiocerol dimycocerosates and phenolic glycolipids among other complex molecules, have evolved with time. These polyketide synthases, may vary in their structure, activities, and genetic recombination may have promoted the emergence of differential virulence factors in MTBC [27]. It is also clear that acquisition of drug-resistance, due to adaptation, restoration of fitness thanks to compensatory mutations, is not always a disabling factor for the spreading of specific emerging multi-drug resistant clones [28–30].

Whatever the evolutionary history and cross-relationships between the pathogen diversity and its hosts, superimposition of various maps suggests that all MTBC lineages did not spread using the same routes. Fig. 6 shows a strange pattern: the superimposition of todays' distribution of L1 and L3 seems to perfectly coincide with specific trade routes: whereas L1 distribution suggests that it has spread via a maritime route, L3 distribution on the contrary suggests that it may have spread by land routes.

TB-Annotator, by its ability to combine SNPs diversity analysis, analysis of known region of differences, discovery of new deletions, gene content and CRISPR, and its automatized workflow management system that allows regular analysis of all publics SRA as well as customized SRA analysis, appears to be a truly innovative and scalable system for genome diversity analysis of the MTBC and reconstruction of its history.

5. Conclusion

Digging into 100,000 MTBC genomes is now feasible thanks to the combination of an increase in publicly available whole-genome sequences of pathogens and to the development of sophisticated bioinformatics integrative and scalable information systems such as *TB-Annotator* v.2. Recent results suggest that we are just at the start of being able to rebuild an exhaustive and deep, global and local, evolutionary history of the MTBC complex thanks to the use of big datasets and machine-learning-based approaches.

The links between animal-TB and human-TB have certainly evolved through time. We are now 8 billions of humans whereas the natural fauna is shrinking and even collapsing in terms of biodiversity. The contemporary transmission of tuberculosis is likely to be very different from the early stages of tuberculosis spreading, that may have been very slow during Paleolithic or even during Neolithic times for obvious demographic reasons.

Moreover, bottlenecks in global history due to climate change or sudden crisis, such as the rapid extinction of the bronze age, or the plague epidemic, are likely to have introduced demographic breakpoints and difficulties to reconstruct the global tuberculosis history [31,32]. It is likely that deciphering the changes in pathogen diversity and virulence across times also poses a lot of challenges for ancient pathogens such as TB, that may have been co-evolving with human beings long before cattle domestication [22], and the global story on host-pathogen interactions is different depending on other parameters including human life-style evolution, and even long-term climate changes. Such a dynamic change in the *Pathocenosis* puts Tuberculosis as a forefront model in terms of host-pathogen cross-talks, evolutionary population

genetics and anthropological studies [33].

Funding

This study did not receive specific funding. MRS was a PhD fellow of the Nigerian PTDF (Petroleum Technology Development Fund).

Author contributions section

GS wrote the "TB-Annotator" information pipeline, with CG supervision, MRS, KL, TBP, JM, FM, FM, ABD contributed to WGS support and analysis, EC and EC contributed to funding support, GS, GR, CG, CS contributed to the analysis of data, GS, GR, CG and CS contributed to first draft writing, CS wrote the revision.

Ethical approval

Not required.

Declaration of competing interest

None declared.

Acknowledgements

We are grateful to Pr. Erik Denamur, Head of UMR1137 for having wellcome us (MRS, CS) in the IAME laboratory, INSERM-UMR1137. All computations presented in this study have been performed on the "*Mésocentre de Calcul de Franche-Comté supercomputer facilities*".

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] Gagneux S. Ecology and evolution of *Mycobacterium tuberculosis*. Nat Rev Microbiol 2018;16:202–13. https://doi.org/10.1038/nrmicro.2018.8.
[2] (NCBI) NCfBI. MD). National library of medicine (us). national center for biotechnology information; 1988. 1988.
[3] Freschi L, Vargas Jr R, Husain A, Kamal SMM, Skrahina A, Tahseen S, et al. Population structure, biogeography and transmissibility of *Mycobacterium tuberculosis*. Nat Commun 2021;12:6099. https://doi.org/10.1038/s41467-021-26248-1.
[4] Napier G, Campino S, Merid Y, Abebe M, Woldeamanuel Y, Aseffa A, et al. Robust barcoding and identification of *Mycobacterium tuberculosis* lineages for epidemiological and clinical studies. Genome Med 2020;12:114. https://doi.org/10.1186/s13073-020-00817-3.
[5] Thawornwattana Y, Mahasirimongkol S, Yanai H, Myat Win Maung H, Cui Z, Chongsuvivatwong V, et al. Revised nomenclature and SNP barcode for *Mycobacterium tuberculosis* lineage 2. Microb Genom 2021. https://doi.org/10.1099/mgen.0.000697.
[6] Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 2018;34:i884–90. https://doi.org/10.1093/bioinformatics/bty560.
[7] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754–60. https://doi.org/10.1093/bioinformatics/btp324.
[8] Seemann T. Snippy, rapid haploid variant calling and core genome alignment. 2020.
[9] Phelan JE, O'Sullivan DM, Machado D, Ramos J, Oppong YEA, Campino S, O'Grady J, McNerney R, Hibberd ML, Viveiros M, Huggett JF, Clark TG. Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. Genome Med 2019;11:41. https://doi.org/10.1186/s13073-019-0650-x.
[10] Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 2014;30:1312–3. https://doi.org/10.1093/bioinformatics/btu033.
[11] Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics 2019;35:4453–5. https://doi.org/10.1093/bioinformatics/btz305.
[12] Senelle G, Guyeux C, Refrégier G, Sola C. TB-Annotator: a scalable web application that allows in-depth analysis of very large datasets of publicly available *Mycobacterium tuberculosis* complex genomes. in preparation, preprint, https://doi.org/10.1101/2023.06.12.526393; 2023.
[13] Palittapongarnpim P, Ajawatanawong P, Viratyosin W, Smittipat N, Disratthakit A, Mahasirimongkol S, et al. Evidence for host-bacterial Co-evolution via genome sequence analysis of 480 Thai *Mycobacterium tuberculosis* lineage 1 isolates. Sci Rep 2018;8:11597. https://doi.org/10.1038/s41598-018-29986-3.
[14] Guyeux C, Senelle G, Refrégier G, Bretelle-Establet F, Cambau E, Sola C. Connection between two historical tuberculosis outbreak sites in Japan, Honshu, by a new ancestral *Mycobacterium tuberculosis* L2 sublineage. Epidemiol Infect 2022;150:1–10. https://doi.org/10.1017/S0950268822000048.
[15] Shitikov E, Kolchenko S, Mokrousov I, Bespyatykh J, Ischenko D, Ilina E, et al. Evolutionary pathway analysis and unified classification of East Asian lineage of *Mycobacterium tuberculosis*. Sci Rep 2017;7:9227. https://doi.org/10.1038/s41598-017-10018-5.
[16] Mestre O, Luo T, Dos Vultos T, Kremer K, Murray A, Namouchi A, et al. Phylogeny of *Mycobacterium tuberculosis* Beijing strains constructed from polymorphisms in genes involved in DNA replication, recombination and repair. PLoS One 2011;6:e16020.
[17] Keim P, Van Ert MN, Pearson T, Vogler AJ, Huynh LY, Wagner DM. Anthrax molecular epidemiology and forensics: using the appropriate marker for different evolutionary scales. Infect Genet Evol 2004;4:205–13.
[18] Menardo F, Rutaihwa L, Zwyer M, Borrell S, Comas I, Conceição E, et al. Local adaptation in populations of *Mycobacterium tuberculosis* endemic to the Indian Ocean Rim [version 2; peer review: 2 approved]. F1000Research 2021:10. https://doi.org/10.12688/f1000research.28318.1.
[19] Shuaib YA, Utpatel C, Kohl TA, Barilar I, Diricks M, Ashraf N, et al. Origin and global expansion of *Mycobacterium tuberculosis* complex lineage 3. Genes 2022:13. https://doi.org/10.3390/genes13060990.
[20] Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, et al. Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. Nat Genet 2013;45:1176–82. https://doi.org/10.1038/ng.2744.
[21] Achtman M. How old are bacterial pathogens? Proceedings Biological sciences/The Royal Society 2016:283. https://doi.org/10.1098/rspb.2016.0990.
[22] Kerner G, Laval G, Patin E, Boisson-Dupuis S, Abel L, Casanova JL, Quintana-Murci L. Human ancient DNA analyses reveal the high burden of tuberculosis in Europeans over the last 2,000 years. Am J Hum Genet 2021;108:517–24. https://doi.org/10.1016/j.ajhg.2021.02.009.
[23] Click ES, Winston CA, Oeltmann JE, Moonan PK, Mac Kenzie WR. Association between *Mycobacterium tuberculosis* lineage and time to sputum culture conversion. Int J Tubercul Lung Dis 2013;17:878–84. https://doi.org/10.5588/ijtld.12.0732.
[24] Khandkar C, Harrington Z, Jelfs PJ, Sintchenko V, Dobler CC. Epidemiology of peripheral lymph node tuberculosis and genotyping of *M. tuberculosis* strains: a case-control study. PLoS One 2015;10:e0132400. https://doi.org/10.1371/journal.pone.0132400.
[25] Billard-Pomares T, Marin J, Quagliaro P, Mechai F, Walewski V, Dziri S, et al. Use of whole-genome sequencing to explore *Mycobacterium tuberculosis* complex circulating in a hotspot department in France. Microorganisms 2022;10. https://doi.org/10.3390/microorganisms10081586.
[26] Cui Z, Liu J, Chang Y, Lin D, Luo D, Ou J, et al. Interaction analysis of *Mycobacterium tuberculosis* between the host environment and highly mutated genes from population genetic structure comparison. Medicine (Baltim) 2021;100:e27125. https://doi.org/10.1097/MD.0000000000027125.
[27] Boritsch EC, Frigui W, Cascioferro A, Malaga W, Etienne G, Laval F, et al. pks5-recombination-mediated surface remodelling in *Mycobacterium tuberculosis* emergence. Nat Microbiol 2016;1:15019. https://doi.org/10.1038/nmicrobiol.2015.19.
[28] Gagneux S, DeRiemer K, Van T, Kato-Maeda M, de Jong BC, Narayanan S, et al. Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. Proc Natl Acad Sci U S A 2006;103:2869–73.
[29] Merker M, Blin C, Mona S, Duforet-Frebourg N, Lecher S, Willery E, et al. Evolutionary history and global spread of the *Mycobacterium tuberculosis* Beijing lineage. Nat Genet 2015;47:242–9. https://doi.org/10.1038/ng.3195.
[30] Merker M, Rasigade JP, Barbier M, Cox H, Feuerriegel S, Kohl TA, et al. Transcontinental spread and evolution of *Mycobacterium tuberculosis* W148 European/Russian clade toward extensively drug resistant tuberculosis. Nat Commun 2022;13:5105. https://doi.org/10.1038/s41467-022-32455-1.
[31] Drake BL. The influence of climatic change on the late bronze age collapse and the Greek dark ages. J Archaeol Sci 2012;39:1862–70.
[32] Gage KL, Kosoy MY. Natural history of plague: perspectives from more than a century of research. Annu Rev Entomol 2005;50:505–28.
[33] Grmek M. Les maladies à l'aube de la civilisation occidentale. Paris: Payot; 1994.
[34] Couvin D, Reynaud Y, Rastogi N. Two tales: worldwide distribution of Central Asian (CAS) versus ancestral East-African Indian (EAI) lineages of *Mycobacterium tuberculosis* underlines a remarkable cleavage for phylogeographical, epidemiological and demographical characteristics. PLoS One 2019;14:e0219706. https://doi.org/10.1371/journal.pone.0219706.
[35] O'Neill MB, Shockey A, Zarley A, Aylward W, Eldholm V, Kitchen A, et al. Lineage specific histories of *Mycobacterium tuberculosis* dispersal in Africa and Eurasia. Mol Ecol 2019;28:3241–56. https://doi.org/10.1111/mec.15120.