An updated evolutionary history and taxonomy of *Mycobacterium tuberculosis* lineage 5, also called *M. africanum*

1.1 Author names

Muhammed Rabiu Sahal[2,3*], Gaetan Senelle[1,2*], Kevin La[4], Barbara Molina-Moya[5], Jose Dominguez[5], Tukur Panda[6], Emmanuelle Cambau[2,4], Guislaine Refregier[7], Christophe Sola[2,3$] and Christophe Guyeux[1]

1.2 Affiliation(s)

[1]FEMTO-ST Institute, UMR 6174 CNRS, DISC Computer Science Department, Univ. Bourgogne Franche-Comté (UBFC), 16 Route de Gray, 25000 Besançon

[2]INSERM-Université Paris-Cité, IAME Laboratory, UMR1137

[3]Université Paris-Saclay, 91190, Saint-Aubin, France

[3]INSERM-Université Paris-Cité, IAME Laboratory, UMR1137

[4]AP-HP, GHU Nord site Bichat, Service de mycobactériologie spécialisée et de référence

[5]Servei de Microbiologia, Hospital Universitari Germans Trias i Pujol, Institut d'Investigació Germans Trias i Pujol, Universitat Autònoma de Barcelona, Badalona, Spain CIBER Enfermedades Respiratorias (CIBERES), Instituto de Salud Carlos III, Madrid, Spain.

[6] KNCV Tuberculosis Foundation, Nigeria

[7] Ecologie Systematique Evolution, Université Paris-Saclay, CNRS, AgroParisTech, UMR ESE, 91405, Orsay, France

1.3 Corresponding author and email address

[$]Christophe Sola, [2]Université Paris-Saclay, 91190, Saint-Aubin, France

[*]These two authors contributed equally to the study

1.4 Keywords

Mycobacterium tuberculosis complex, L5, Mycobacterium africanum, Evolutionary Genomics, Phylogeography

1.5 Repositories:

The search was done in the TB-Annotator 15901 genomes version which is available at: http://51.178.57.127:8082/ (to be added). The new sequenced genomes are available via NCBI Bioproject accession number: (to be added)

Abstract

Contrarily to other lineages such as L2 and L4, there are still scarce whole-genome-sequence data on L5-L6 MTBC clinical isolates in public genomes repositories. Recent results suggest a high complexity of L5 history in Africa. It is of importance for an adequate assessment of TB infection in Africa, that is still related to the presence of L5-L6 MTBC strains. This study reports a significant improvement of our knowledge of L5 diversity, phylogeographical history, and global population structure of *Mycobacterium africanum* L5. To achieve this aim, we sequenced new clinical isolates from Northern Nigeria and from proprietary collections, and used a new powerful bioinformatical pipeline, *TB-Annotator* that explores not only the shared SNPs but also shared missing genes, identical IS*6110* insertion sites and shared regions of deletion. This study using both newly sequenced genomes and available public genomes allows to describe new L5 sublineages. We report that the MTBC L5 tree is made-up of at least 12 sublineages from which 6 are new descriptions. We confront our new classification to the most recent published one and suggest new naming for the discovered sublineages. Finally, we discuss the phylogeographical specificity of sublineages 5.1 and sublineage 5.2 and suggest a new hypothesis of L5-L6 emergence in Africa.

Impact statement

Recent studies on *Mycobacterium africanum* (L5-L6-L9 of MTBC) genomic diversity and its evolution in Africa discovered three new lineages of the *Mycobacterium tuberculosis* complex (MTBC) in the last ten years (L7, L8, L9). These discoveries are symptomatic of the delay in characterizing the diversity of the MTBC on the African

continent. Another understudied part of MTBC diversity is the intra-lineage diversity of L5 and L6. This study unravels an hidden diversity of L5 in Africa and provides a more exhaustive description of specific genetic features of each sublineage by using a proprietary "*TB-Annotator*" pipeline. Furthermore, we identify different phylogeographical localization trends between L5.1 and L5.2, suggesting different histories. Our results suggest that a better understanding of the spatiotemporal dynamics of MTBC in Africa absolutely requires a large sampling effort and powerful tools to dig into the retrieved diversity.

Data summary

[A section describing all supporting external data, software or code, including the DOI(s) and/or accession numbers(s), and the associated URL. If no data was generated or reused in the research, please state this.]

The search was done in the TB-Annotator 15901 genomes version which is available at: http://51.178.57.127:8082/ (to be added). The new sequenced genomes are available via NCBI Bioproject accession number: (*to be added*). The authors confirm all supporting data, code and protocols have been provided within the article or through supplementary data files**.**

Introduction

Mycobacterium africanum is a human and animal pathogenic species that belongs to the *Mycobacterium tuberculosis* complex (MTBC) and that was identified for the first time in 1968 (Castets et al., 1968). It was rapidly shown to present intermediate phenotypic characteristics in between human and bovis tubercle bacilli (David et al., 1978), and later split into two highly related phylogenetic branches, L5 and L6 with specific geographical distribution (Comas et al., 2013). L5 is also known as *M. africanum* West african 1 (MAF1) and as AFRI2-AFRI3 regarding its spoligotype pattern, whereas L6 is known as *M. africanum* West african

2 (MAF2) and as AFRI1 (Comas et al., 2009). L5 is more prevalent in the gulf of Guinea (Ghana, Benin, Nigeria), whereas L6 is more prevalent in the Western part of West Africa (Senegal, Gambia) (de Jong et al., 2009; Gehre et al., 2013). Recently, two important papers pointed the undersampling of L5 and L6 in public MTBC genomes database and laid the foundation for a first sub-classification and diversification history (Ates et al., 2018 ; Coscolla et al., 2021). It is hypothesized that L5 and L6 could have been endemic to different specific african populations before european colonization, and that these genetic variants could have been less pathogenic than the clinical isolates of the L4 lineage, due to a longer history of co-evolution with humans. Specific L4 sublineages are today highly prevalent in Africa, e.g. L4.6.1 ”Uganda”, L4.6.2 ”Cameroun”, L4.7 ”Congo”, and other modern strains such as L4.8 (Asante-Poku et al., 2015; Malm et al., 2017; Niobe-Eyangoh et al., 2003; Wamala et al., 2015). Hence, the differential history of L4-L5-L6 Africa is of great importance to study the global history of TB. A recent national sampling of MTBC positive Ziehl-Neelsen slides in Nigeria pointed to a very high prevalence of L5 in the south/south-east states of Nigeria (Molina-Moya et al., 2018b). This dense region is geographically close to the cradle of agriculture in Africa, around South-east Nigeria and Cameroon (Molina-Moya et al., 2018b; Phillipson, 1975). It is widely accepted that, once agriculture emerged, the feeding of animals was the first goal of early African growers to maintain foraging and a nomadic way of life. In Africa, and contrarily to the rest of the world, cattle predominated over crops in the food production process. This might have given more time and more chances for animal diseases to spread into human populations, in relation to slow the known environmental changes, such as the ones studied in North-East Africa, and to a slow demographic expansion, slowly accompanied by the Bantu agricultural expansion (Kuper and Kropelin, 2006).

Three main issues are found when dealing with L5 phylogenetics. The first one is undoubtedly the poor sampling and hence the poor statistical representativity of the population studied (Ates et al., 2018). The second one is the important geographical bias that is found in data content in databases, since some

4

demographically important countries and populations, e.g. Nigeria, are only weakly characterized whereas less populated countries, yet more developed, such as Ghana for example, or more recently Ethiopia, start to be well characterized. The third issue is the potential biases and loss of information in evolutionary studies when genome characterization is inferred solely using H37Rv as the reference genome (Sanoussi et al., 2021).

Nevertheless, even with these limitations, some important results were obtained recently. Firstly it is suggested that L5 and L6 emerged in two different ecological niches (Asante-Poku et al., 2016; Comas et al., 2013; Otchere et al., 2018). Whereas L5 is strictly human, L6 is closer to *M. bovis* (Otchere et al., 2018). Another important result is the finding of a new L9 lineage, that is found in East Africa, and the finding of a complex history of *Mycobacterium africanum* (Coscolla et al., 2021). In this last study, Coscolla *et al*. were not too precise in the definition of some sublineages of L5, for example by directly jumping to sub-sublineages of L5.1, thus showing their difficulties to find adequate markers to define hierarchical levels (Coscolla et al., 2021). In another perspective, Sanoussi *et al.* sequenced three new L5 genomes (from Gambia, Benin and Nigeria) using the zero-mode wave guide technology (PacBio®, i.e. producing long read sequencing), and they performed *de novo* assemblies of these new genomes. They showed that classical genome comparison to H37Rv failed to detect some potentially physiologically important genes, a strong limitation in comparative genomics studies that usually tend to limit their analysis to SNPs variation (Sanoussi et al., 2021). More precisely, the pangenome of L5 and L6 is likely to differ from the one of other MTBC lineages. The gene content variation could be larger than expected initially. In addition, gene expression impacting metabolism and pathogenic power is likely to provide different metabolic specificities in L5 versus other lineages (Sanoussi et al., 2021; Yimer et al., 2020).

For all these reasons, we undertook a new study using Whole-genome Sequencing (WGS) of newly recruited L5 and L6 clinical isolates from an african country,

Nigeria, as well as from archives from previous studies. We analyzed the genome reads using a new proprietary informatic pipeline, *TB-Annotator*, in order to perform an *in-depth* phylogenetical analysis of this larger collection (Senelle *et al*. manuscript in preparation). We report in this article the description of an improved L5 phylogenetic tree with detailed evolutionary events list for each branch, and we suggest new hypotheses regarding its history.

Methods

Origin of clinical isolates of newly sequenced L5-L6 genomes

The 25 newly sequenced *M.africanum* (numbered CUS0000001- CUS0000025) were recovered from well-characterized human TB cases from archive DNA from previous published studies whose DNA had previously been sent between 2007 to 2019 to the Infection Genetics and Evolution of Emerging Pathogens research team (IGEPE, Institute of Integrative Cell Biology, Gif sur Yvette) (Lawson et al., 2012; Molina-Moya et al., 2018a; Molina-Moya et al., 2018b). Another origin of samples is culture positive L5-L6 isolates from the French national TB reference-associated laboratory in Bichat Hospital Paris, France. Altogether, these included isolates recovered from Nigeria (n=13), and France (n=12) with 11 patients representing seven West-African countries nationalities and one isolate representing a patient of French nationality (data available upon request).

The selected samples were initially screened as L5-L6 based on a simple MIRU24 locus copy number assay, and further confirmation by a Line-probe PCR-based assay (Hain Lifescience GmbH-Brücker Molecular diagnostics, Gehren, Germany). CUS0000001-09 and CUS0000022-25 are archive DNA from Nigeria whereas CUS0000010-21 are L5-L6 isolates from Bichat Hopital. CUS0000001 and 0000008 represent isolates from Lawson *et al.* 2012 collection, while the remaining eleven are from collections either recovered from Molina-Moya *et al.*, 2018a or 2018b collections. In this study, newly recruited L5-L6 DNAs from Nigeria were selected from patients with a microbiological diagnosis of TB characterized by the presence of 2 copies of MIRU24 as assessed by MIRU-VNTR analysis. The demographic data (age, sex, country of birth, years of diagnosis especially applicable to France isolates) and clinical (location of disease, sputum smear status, and previous diagnosis of TB) characteristics of the patients, when available will be reported elsewhere.

Culture, DNA purification methods, and Whole-Genome Sequencing on Illumina

MTBC clinical isolates were firstly inactivated in P3 laboratory from Coletsos solid media grown cultures. DNA was extracted using DNeasy® UltraClean 96 Microbial Kit (Qiagen®, Hilden, Germany). A Nanodrop® device (ND-ONE-W) was used to control purity and concentration for each sample (Thermofisher®, Waltham, USA). The Qiagen® kit allows to obtain good quality purified DNA as assessed by a 280/260 and 260/230 ratio close to 2 and the concentration of the DNA was sufficient to perform library preparation (concentration ~100 ng/µl). Library was produced using the Nextera® XT DNA Library Preparation kit (Illumina, San Diego, USA)- with a control of the fragment size on a 2100 Bioanalyzer® Instrument (Thermofisher®, Waltham, USA). Sequencing was performed using the Miniseq® High Output Kit (Illumina®, San Diego, USA) generating paired-end reads of 150 bp. Quality of reads was then analyzed by *fastqc* (Andrews, 2010).

Variant definition and presentation of the *TB-Annotator* and *CRISPR-builder-TB* pipelines

Figure 1 shows the general algorithm that was used to process the Short Read Archives (SRA) produced in house or recovered in public repositories. A brief description of the pipeline can be found in (Guyeux et al., 2022). At the time of writing, the V1 database contained 15909 SRA; results were also checked in part using the expanded V2 version (Senelle *et al*. in preparation).

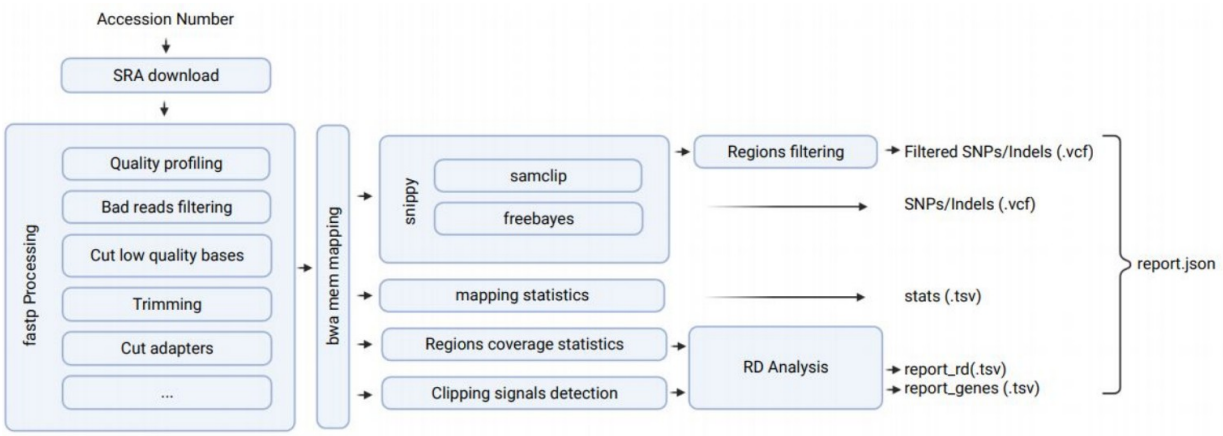


Figure 1: General algorithm

Selection of samples from the Public databases for uploading on TB-annotator platform

NCBI SRA Projects from all studies known or likely to include L5 or L6 samples were incorporated in TB annotator platform using the keywords "*M. africanum* ; *Mycobacterium africanum* ; *Mycobacterium tuberculosis* L5 ; *Mycobacterium tuberculosis* L6".

Sample selection for the study, TB-annotator platform, and tree reconstruction

All samples sharing the defining SNP shared by L5 samples were selected; as L5 defining SNP we chose the C9566T SNP (selection of 360 isolates with 133 exclusive variants and 548 more at >95% specificity) instead of C1799921A (selection of 359 isolates with only one exclusive variants and 676 variants at >95% specificity. They confirmed to be monophyletic in the visualization tool. Several subselections were performed to further explore L5 intra-lineage diversity. Pairwise distances between clusters were systematically computed for the defined clades in order to asses the epidemic (< 15 *SNP*) versus historic (>15 SNP) status. Because of complexity issues, this is done only for clusters of size (< 200 isolates). The unrooted tree was built using a similar approach as the one published in (Guyeux et al., 2022). Briefly, we used RAxML on a multi fasta-file produced using *TB-Annotator*; our files initially

contained 15901 SRAs at start, the bincat model was used instead of the GTR model and no outgroup was chosen (Stamatakis, 2014). A confirmation of the results using the extended *TB-Annotator* v2.1 version 85000 SRA, that unfortunately was not much richer in L5, was obtained (results not shown). Variants were searched in *TB-Annotator* v1 or v2.1 based on SNP position or using the SPDI Model, with H37Rv as the reference genome (NC_000962.3 reference sequence) (Holmes et al., 2020).

Results

3.1 Identification of new L5 sublineages

A first unrooted tree of L5, shown in **Figure 2**, was obtained on an alignment analyzing 360 L5 samples. This first tree was obtained using not only SNPs but also presence of IS*6110* at the same position, shared deletions and shared absence of H37Rv genes. This original combined sample requires to use binary info of shared(1)/unshared(0) patterns.

We first confirm the overall predominance of the L5.1 sublineage in the current sampling as described by Coscolla et al (2021). Second, we confirm the existence of other L5 sublineages such as L5.2 and L5.3 (Coscolla et al., 2021).We identified 6 new subclades, first identified as 5.a to 5.f. Some of them grouped with L5.1 main sublineage, namely those transiently named 5.c, 5.d and 5.f. The subgroup including them and former 5.1.1, 5.1.2 and 5.1.3, supported by 30 SNPs, is denoted G2 below. The third main monophyletic group named G3 consists of subclades 5.1.4 and 5.1.5, supported by 126 SNP, and the last main monophyletic group, G4, supported by 72 SNPs, includes 5.2, 5.3 and two new subclades 5.a and 5.b.The new clade transiently denoted 5.e, at the bottom of the tree on **Figure 2** (also noted G1 thereafter), seems basal but this is due to the fact that the tree was unrooted (see below for a description of its relative position in L5 diversity).

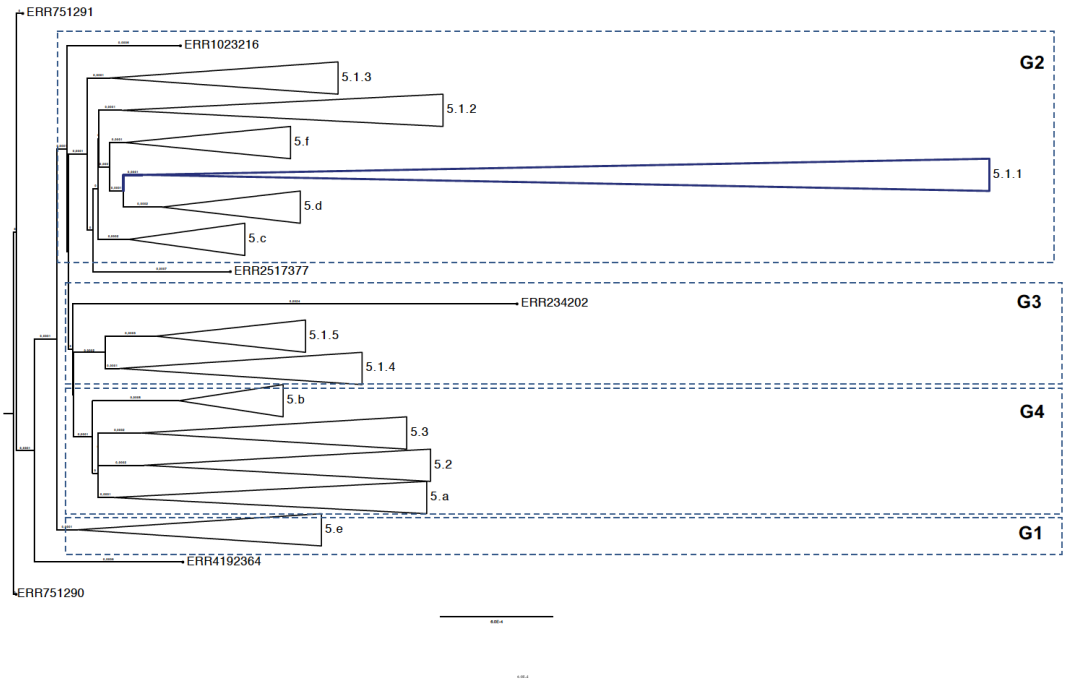


Figure 2: Unrooted tree of L5 samples defining main phylogenetic groups

On the one hand, 5.e seems to have a basal position. On the other hand, when we look at the arrangement of the 3 other groups, it appears as shown in Figure 2 that 5.1.4. and 5.1.5 do not clade with 5.1.1, 5.1.2 and 5.1.3, but rather with 5.2 and 5.3, which would argue against the current nomenclature and scheme proposed in (Coscolla et al., 2021). However, the position of the ancestral nodes of these three groups is rather confusing, and this is reinforced by the weakness of the supports of the branches near the root (data not shown). The phylogeny does not seem to be resolved in view of the very short length of the branches there: as often in such cases, the further back in the tree one goes, the more the phylogenetic relationship becomes blurred. Before going into more detail on the new clades and sublineages definition, we will first investigate this lack of resolution at the base of the tree, and then propose a new phylogeny of L5.

3.2 Genetic features characterizing the main L5 large monophyletic groups

We analyzed the G1 to G4 groups by pair and by triplet, to see which tuples had similar mutations, a sign of a common ancestor. G1 and G2 have 4 exclusive SNP variants at 97.64%, meaning that all the strains of this pair have these 4 mutations (C558750T, G4313357A, T1407273A, and G4343653A), while we can find 5 strains outside with these mutations. Conversely, the other pairs of strains shared nothing in common. At the triplet level, (G1, G2, G3) shared 10 exclusive SNP variants, and 19 more SNP variants with exclusivity being larger than 95%, while no other triplet share any other exclusive element.

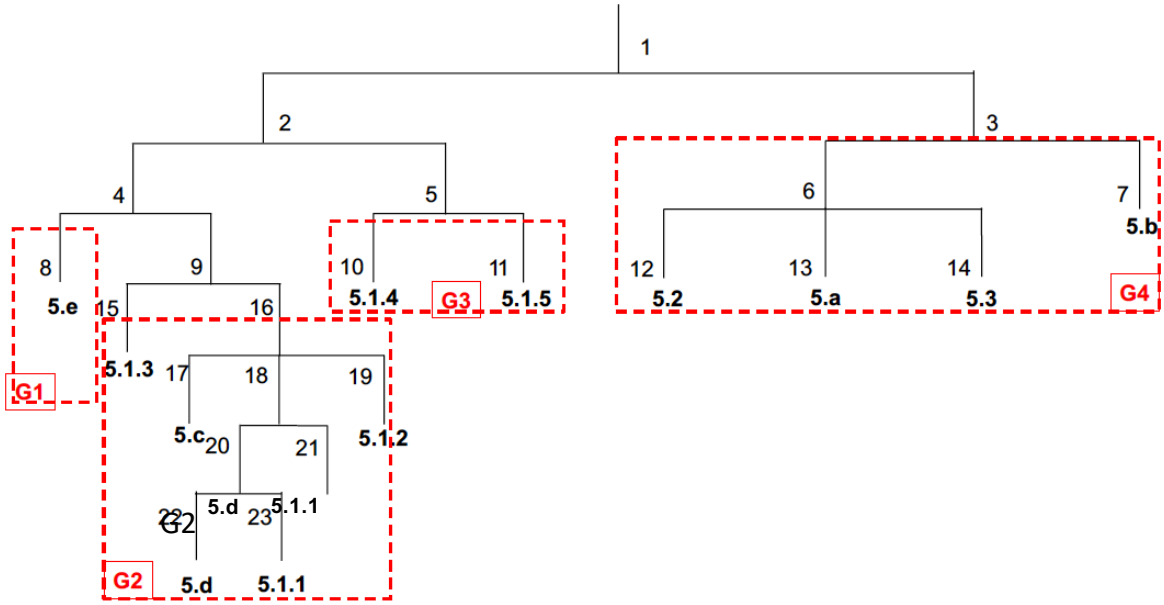


Figure 3: Consensus phylogenetic relationships between reference sublineage naming (Coscolla et al., 2021) and transient naming of new clades (this study) (new 5.3 not present at this stage)

We took into account that only (G1, G2) on the one hand, and (G1, G2, G3) on the other hand, emerge from our exclusivity study to build the tree detailed in **Figure 3** based on **Figure 2**. **Figure 4** presents the same tree, however following a hierarchical naming process.

In this phylogeny proposal, the only internal node not having its position proven through the data is the node giving rise to L5.2, whose resolution may still be problematic. However, we will see in the CRISPR locus structural study section of this article that a recombination of IS*6110* upstream of spacer 35 is a specific feature of 5.1, which is never present in 5.2, and this fact strengthens our proposal. Moreover, this tree is in good accordance with that of Coscolla *et al.* (2021). In the following, we will present the elements supporting the different nodes of the proposed tree.

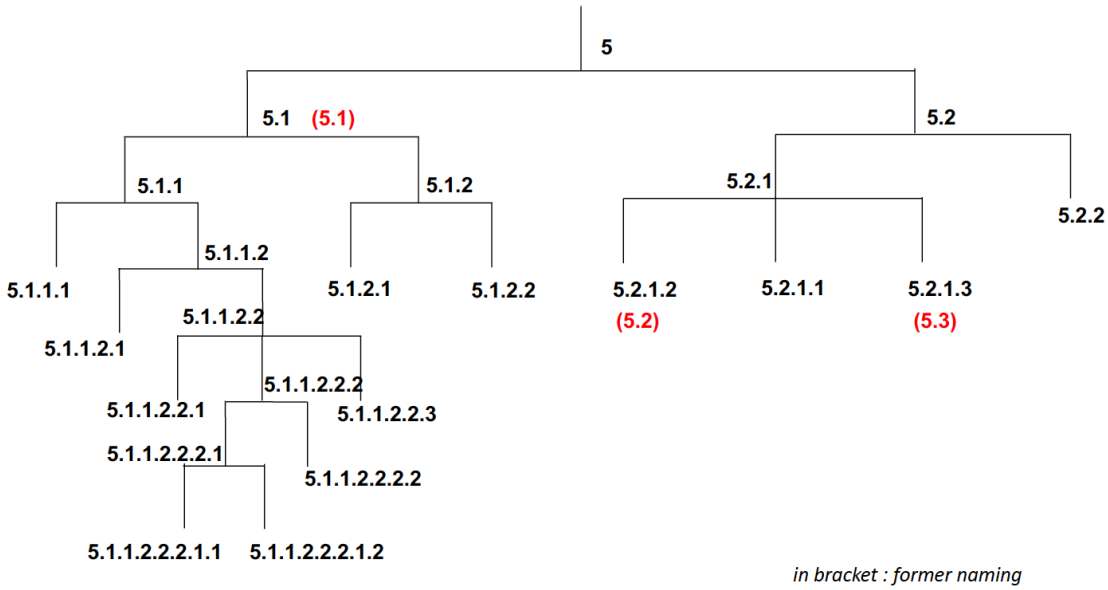


Figure 4: Consensus phylogenetic relationships with new hierarchical level naming and reminders of main correspondence with M. Coscolla *et al.* classification (under brackets) (new 5.3 is not present at this stage)

3.3 Genetic features characteristic of the main sublineages

Our first objective is to provide support for our proposed L5 tree, as depicted in **Figure 4**. We describe below the specificity of each node (internal or leaf) starting from its root, on a sample of 359 and not 360 since ERR4872200 does not contain any of the 5.1, 5.2, and 5.3 defining SNPs. The classification of this SRA remains pending. The L5 lineage has 133 exclusive SNP variants of its own, and 548 variants with more than 95 % exclusivity, (cf. **Supplementary Material 2**). A SNP in *gyrA* for example, very well defines L5, namely C9566T (n=360). The absence of three specific L5 genes, Rv2478c, Rv1978, and Rv1979c also defines L5. Two isolates, ERR439931 and ERR4872094 over 360 did not lack Rv2478c. Alternatively the lack of Rv2478c is found in >360 isolates in the *TB-Annotator* database, among which ancestral L2 from Kwa-Zulu Natal (SRR10315503). Concerning the known region of differences, RD713 (Rv1977-Rv1979c) is common to all L5 (Mostowy et al., 2004). A search of lack of Rv1978 in *TB-Annotator* retrieves 390 SRAs, among which a few L2, L3, L6 and animal isolates (Mostowy et al., 2004). Last but not least, 350 strains (97.49% of L5) have an IS*6110* at position 4185790, while only 10 out of 15542 not-L5 isolates distributed throughout the tree have this IS too. To put it in a nutshell, 133 SNPs define this lineage (with

548 exclusive >95% variants), these other characters being listed as almost exclusive. Let us go further into the tree.

First sublevel.

Internal node 5.1 is defined by T3452C, and is constituted by 278 isolates. In addition to its 10 exclusive SNP variants (plus 19 at > 95% exclusivity), we can note the systematic absence of 3 genes (Rv1333, Rv1334, and Rv1335), the first two being almost exclusive, corresponding to RD711 as region of difference (10 isolates too outside this selection) (Mostowy et al., 2004). Section 3.5 on CRISPR will provide another characteristic of this lineage. Of note, in this L5.1 as well as in L5.2 and L5.3 nodes, we identify more reliable subgroups than in the previous nomenclature. This requires a renaming of all groups, which is detailed below and with a correspondence with former publication as shown in **Table 1**.

Table 1 : conversion table between the reference nomenclature and our new naming

Coscolla et al. 2021 naming	This study
5.1.1	5.1.1.2.2.3.2
5.1.2	5.1.1.2.2.2
5.1.3	5.1.1.2.1
5.1.4	5.1.2.1
5.1.5	5.1.2.2
5.2	5.2.2
5.3	5.2.3

The node 5.2 encompasses the lineages 5.2.2, 5.2.3 (Coscolla et al., 2021), and two new discovered ones, namely: 5.2.1 and 5.2.4 (**Table 1 and Figure 4**). The 75 strains of L5.2 are defined by exactly one exclusive variant, namely C3471147T (0 at 95%). They have an average of 578 SNPs of difference (standard deviation of 150).

Our last analyses suggest that the L5 tree should include another branch that we propose to name L5.3 (that differs from the former L5.3 from Coscolla *et al*) . This new leaf 5.3 completes this first level of the phylogenetic tree. It is constituted by only 6 isolates at present, but could represent an undersampled subbranch of the tree

(cf. **Figure 5 and Figure 10**). The 6 known samples share 104 exclusive variants (0 at 95%), one of them being C96339T. Their mean distance is equal to 219, with a standard deviation of 95. This clade was not described in the literature nor in *TB-Profiler* (Napier et al., 2020). Three of these genomes had been previously studied in the largest study performed until now on L5-L6 genomes in Ghana and published in (Otchere et al., 2018).

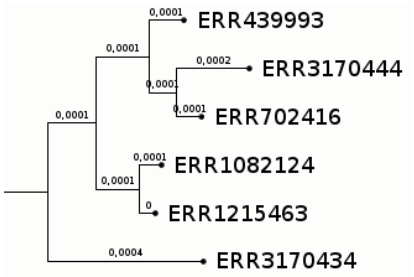


Figure 5: 5.3 sublineage

Second sublevel, with 5.1 as ancestor.

The node 5.1.1 (212 isolates) is the ancestor of lineages 5.1.1.2.2.3.2 (5.1.1), 5.1.1.2.2.4 (5.1.2), and 5.1.1.2.1 (5.1.3), and of a collection of new sublineages (all number in brackets referring to the classification by (Coscolla et al., 2021). This clade, can be defined by the SNP C558750T. It has 2 exclusive SNP variants, and 3 other at 95%. Rv1523 is absent in all these strains, but this is not an exclusive character, as 38 clinical isolates outside the clade share the same deletion.

The node 5.1.2 (65 isolates) is the ancestor of 5.1.2.1 (5.1.4) and 5.1.2.2 (5.1.5) (again all number in brackets referring to (Coscolla et al., 2021)). L5.1.2 is defined by the SNP C118646T and 38 SNP variants are exclusive (and 6 others at 95%). It is constituted by 65 isolates, which are at an average distance of 417 SNPs (standard deviation of 113).

Second sublevel, with 5.2 as ancestor.

L5.2.1, at the top of the tree in **Figure 6**, is constituted by two isolates only: ERR439960 and ERR439985, both from Benin. These two isolates are far from each other (108 SNPs of difference), and far from all other strains, with 152 exclusive variants (0 > 95%), one of them being C15558A. Furthermore, they share an IS*6110* at position 3841699, and a region of difference at position 3738497-

3738697 (200 bp).

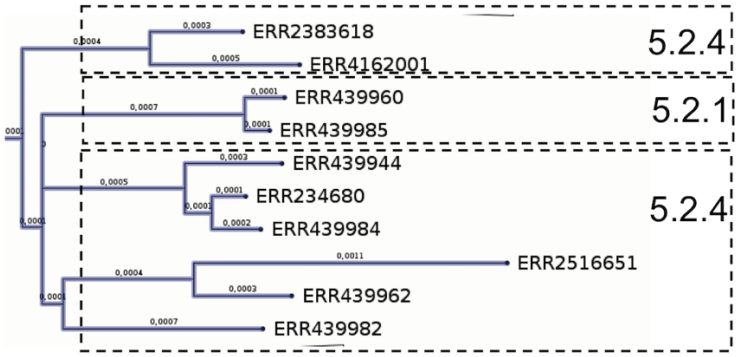


Figure 6: 5.2.1 and 5.2.4 sublineages

L5.2.2 (5.2 in in (Coscolla et al., 2021)) is constituted by the 36 isolates of **Figure 7** that are defined by G30475A (63 exclusive variants, 0 at 95%). 35 isolates have an exclusive IS*6110* at position 2129642, which is missing in ERR2704812 originating from Cameroon. The mean distance is of 438 SNPs, and the standard deviation is equal to 162.

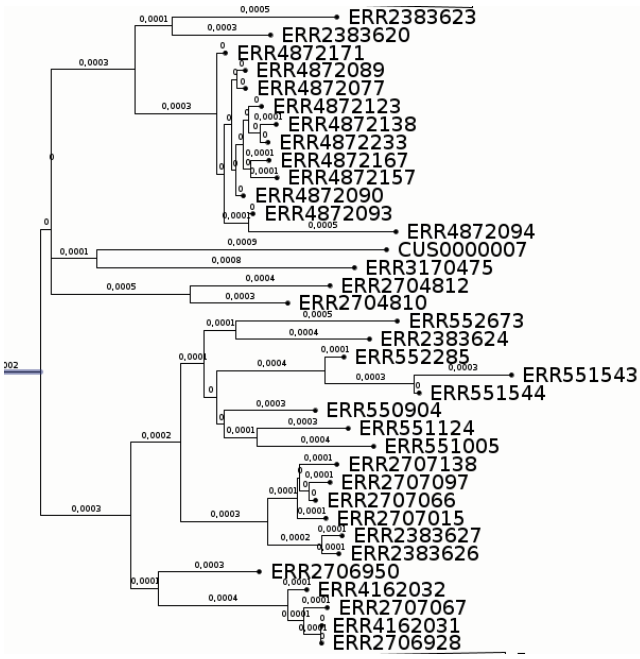


Figure 7: L5.2.2 sublineage or L5.2 in (Coscolla *et al.*, 2021)

L5.2.3 (5.3 in (Coscolla et al., 2021)) has 29 isolates defined by A47099T, as shown in **Figure 8.** They share 59 exclusive variants (8 > 95%) and present an average distance of 432 SNPs (standard deviation: 117). L5.2.4, the remainder strains of

Figure 6, is made-up of 8 isolates. They have two exclusive variants (e.g., G587327A) and a mean difference of 514 SNPs (standard deviation: 124).

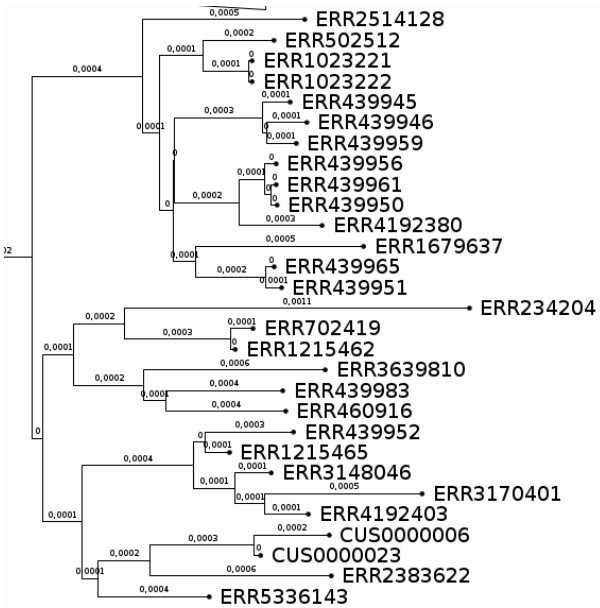


Figure 8: 5.2.3 sublineage

3.4 A thorough classification of the L5 sub-lineages

Further details of all phylogenetic trees built for L5.1 and L5.2 sublineages are described in detail in the **Supplementary Material 2.** The findings on new phylogenetically relevant SNPs, on SNP distances inside these phylogenetical branches, on SNPs specificity and on CRISPR locus architecture can else be found in the **Supplementary Material 1-4**.

3.5 CRISPR locus structure investigation

The CRISPR locus has long be known as a locus evolving by deletion and with good global concordance with SNP-based phylogenies. We describe below its structure in the different L5 sublineages. In what follows, spacer numbers are provided in the extended nomenclature at 68 spacers, whereas SIT there are given to provide the spoligotype value with former 43 spacers format (Couvin et al., 2019; Kamerbeek et al., 1997; van Embden et al., 2000). Lineage 5 is defined by the absence of spacers 18-22 on the one hand, and 51-60 on the other in the 68 spacer format (Brudey et al., 2004); (Refrégier et al., 2020). They all have IS*6110* located upstream of spacer 35. They do not have specific spacer mutants. Unlike the strains of L2, L3 and L4, they do not have DR6 between spacers 65 and 66, but have DR4

135 and DR5 between spacers 66-67, and 67-68 respectively (Refrégier et al., 2020). In
136 terms of duplications, there is only that of spacer 35 between spacers 41 and 42 to
137 note. And as we will see below, recombinations between IS*6110* are partly
138 responsible for the diversity of spoligotypes in the lineage. As stated previously, all
139 L5.1 isolates have lost their 4 spacers in front of spacer 35 (Ates et al., 2018). It
140 sometimes occurs in L5.1 that the deletion is a bit larger than these 4 spacers.
141 Conversely, these 4 spacers or at least one of them are almost always present in
142 L5.2 and L5.3 (Ates et al., 2018). A further investigation of this specific CRISPR
143 area shows that, during the L5.1 diversification, a recombination between the
144 usual IS*6110* in front of spacer 35, on the one hand, and an IS*6110* after spacer
145 30, on another hand, has led to the appearance of such a hole at the origin of
146 Lineage 5.1 (Guyeux et al., 2021; Refrégier et al., 2020). As it is hypothesized
147 that the CRISPR locus was acquired once only, by lateral gene transfer from an
148 environmental host, acquired spacers are found only in lineages having predated
149 MTBC last common ancestor, and evolve since by deletion only[1], and it is not
150 surprising that this hole is found in all descendant lineages of L5.1. Other points
151 that can be made about CRISPR in Lineage 5 include:

152 • low copies number of IS*6110* in CRISPR, and thus reduced possibilities for evolution
153 in this locus, as in the rest of the genome, this considerably slower the pace of
154 evolution of the CRISPR locus.

155 • few modifications in the regions bordering the CRISPR (the CRISPR-associated
156 genes *cas* on the left, and the part flanking the locus on the right).

157 All the characteristics of the CRISPR locus that have produced using "*CRISPR-builder-*
158 *TB*", or produced automatically, or collected from other sources (London School of
159 Tropical Medicine, Jody Phelan, personal communication) been gathered for all L5
160 studied isolates can be found in **Supplementary Material 3.**

161

162 ### 3.6 SNP discovery and Specificity

[1] In the same vein, the more one advances in the lineage, the more the spoligotype is in confetis.

163 When comparing the SNPs informativity level found in this study and the one described in the
164 supplementary material of Coscolla et al. 2021 (n=312), we found an excellent agreement. For
165 some sublineages such as the former L5.1.1 (L5.1.1.2.2.3.2 in this study), the agreement was
166 low (42%), whereas for other subbranches such as former L5.1.3 (L5.1.1.2.1 in this study) the
167 agreement is very high (94% in this case). This is likely to be due to the fact that, given the
168 undersampling of our global dataset, true SNPs discovery is less supported than false SNP
169 discovery for sublineages that are presenting a higher, yet only very poorly investigated,
170 genetic diversity (cf. **Supplementary Material 4**). Since the search for SNPs in *TB-Annotator*
171 was done using a hierarchical process (SNP position followed sometimes by SPDI search), we
172 sometimes unraveled interesting genetic variants. One such example is the case of the finding
173 of three variant alleles in various lineages at position 2187276 (L5.2.3 A:G, L4.1.2 A:T and
174 another L4.1 A:C). Each SNP specificity was verified by individual search in the *TB-Annotator*
175 v1 database and pipeline (see **Supplementary Material 4**). Three other likely convergent but
176 interesting variants detected by Coscolla *et al.* our by this study are
177 SPDI:NC_000962.3:1738750:C:A in *ileS* that is found in both L5.2.2 and in two L3
178 sublineages, SPDI:NC_000962.3:2361310:C:T in *helZ* that is found in L5.1.12.2.2 and in all
179 L4.5 genomes, and SPDI:NC_000962.3:2433085:G:A in *lppM* that is found in L5.2.2 and in
180 the L2 sublineage L2.2 M5.

181 Our pipeline is analyzing all genes, without discarding PE-PPE genes nor other repetitive
182 sequences such as Insertions Sequences, however only 82 mutations were found in PE-PPE
183 genes (**Supplementary Material 4**).

184 To confirm the phylogenetic relationships obtained from binary analyses, we extracted
185 the SNP from 51 samples representative of the L5 diversity (random selection). The
186 corresponding alignement totalled 8701 nucleotides including those belonging to PE-PPE. We
187 built a tree implementing a standard molecular evolution model. We retrieved the same
188 topology as with binary data except for some subbranches of L5.2.4, some of which clustering
189 with L5.2.2 (**Supplementary Material 5**). To confirm that this slight difference did not
190 originate from including PE-PPE genes, we built a second alignment excluding PE-PPE SNPs
191 (**Supplementary Material 5**). The 7840 nucleotides alignment resulted in the same phylogeny
192 except for L5.2.4 subbranches, with a mean 5% increase in branch lengths. This confirms that
193 the proposed classification is highly reliable, with the exception of L5.2.4. This also indicates
194 that PE-PPE SNPs contribute to around 10% of the retrieved diversity and that this diversity
195 adds only little noise to the rest of the phylogenetic information.

3.7 Geographical specificity of L5.1 versus L5.2

We tried to search for an eventual geographical specificity of L5.1 and L5.2 by trying to link the taxonomic position of the main sublineages described to their known geographical region of origin (**Figure 9** and **Supplementary Material 3**). We adopted the same country classification as the one found in (Coscolla et al., 2021). West of West Africa includes Gambia, Senegal, Mauritania, Sierra Leone, Liberia, Guinea, Ivory Coast and Mali; East of West Africa includes Ghana, Benin, Niger, Nigeria and Burkina Faso; Central Africa includes Cameroon, Gabon, Congo, Democratic republic of Congo, and Equatorial Guinea. By performing a Fisher exact statistical test to look for the homogeneity of distribution of sublineage 5.1 and L5.2, we suggest a statistical trend that L5.1 and L5.2 are heterogeneously distributed in the three african regions (p-value = 0.07439, close to 0,05). To support this, we provide a map of the distribution of L5.1 and L5.2 in **Figure 9**. We suggest that L5.1 and L5.2 sublineages have different phylogeographic origin and different epidemic histories, as what had already been demonstrated for L6 but not for L5 in (Coscolla et al., 2021). However this result needs to be confirmed in a new study on an more representative and extended dataset.

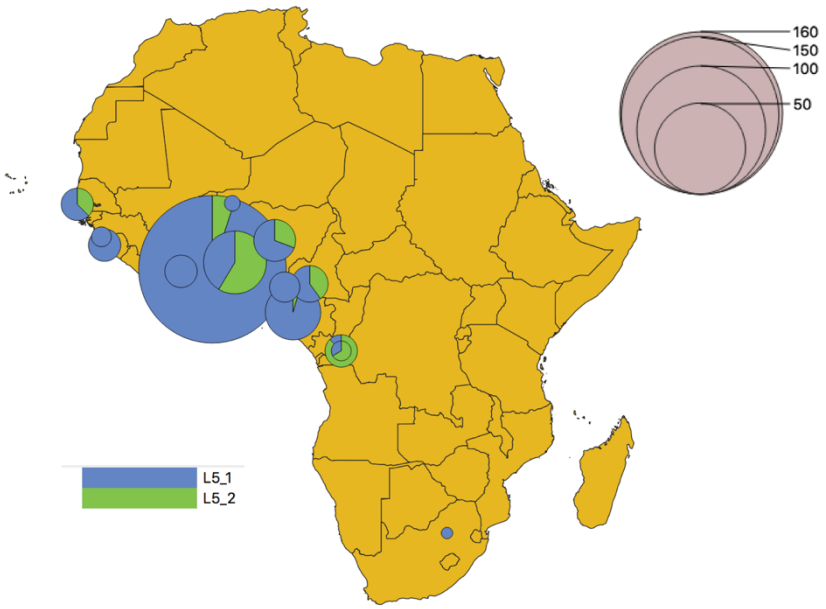


Figure 9: L5.1 and L5.2 (newly defined) geographical distribution

4. Discussion

The reconstruction of the evolutionary history and taxonomy of L5 is important for the understanding of the natural history of tuberculosis in Africa and to predicts its evolution. It is a yet fragmented endeavor since there remains a strong under-representativity of these genomes in public databases. However, as shown in **Figure 10,** that summarizes our results, this study provides a considerable increase in the phylogenetical analysis of L5: (1) a better description of basal groups, (2) the identification of new reliable SNPs at each subbranch (3) and an integrated phylogenetic reconstruction of the entire L5 lineage.

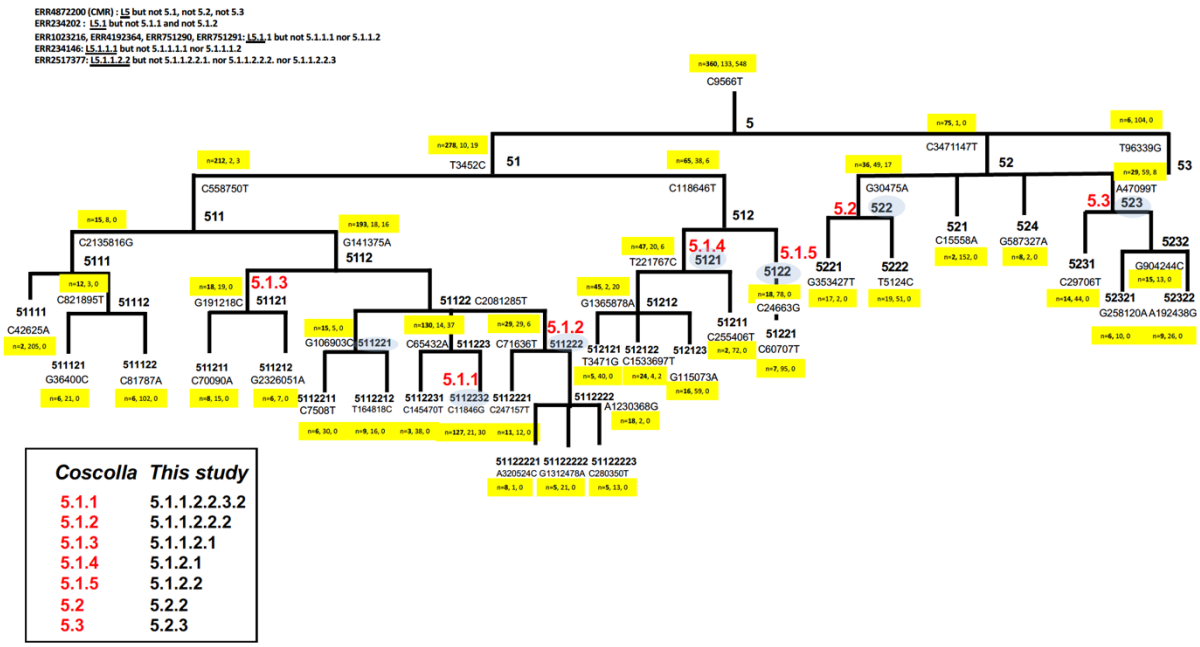


Figure 10: A final phylogeny of Lineage 5 with SNP barcode proposal. *in red : ancient naming (Coscolla et al. 2021); yellow frame: number of genomes in this group, number of defining SNPs (100% exclusive), number of SNPs (>95% exclusive)*; *top left: peculiar isolates IDs that do not fit within this classification*

Our findings comfort the recent studies found on L5, namely the predominance, in the present sampling of what we still call L5.1 (Ates et al., 2018; Coscolla et al., 2021).

The main aim of this study was to build an exhaustive collection of representative L5 genomes and to analyze their structure using a combination of bioinformatical analytical pipelines, so called "*TB-Annotator*". First, let us note that the phylogenetic resolution is more difficult towards the root, which is

236 expected. However it can be solved with the so-called exclusive variants
237 approach. For example, we were able to cleanly define lineage 5.1 (new and old,
238 no modification), which was previously undefined (Coscolla et al., 2021). At the
239 leaf level, our study goes much further than what was currently known with a
240 finer hierarchical classification and identification of new signatures SNPs.

241 A first notice that jumps out is that the tree is that internal nodes are numerous,
242 relatively wide as compared to its depth. This relates with the fact that the means and
243 standard deviation of the differences by sublineage are often large and similar
244 between them, and that the part at > 95% is often null or low, while the number of
245 shared variants is frequently high. In our opinion, these various remarks suggest that
246 diversification *i.e.* transmission was pervasive in the beginning of L5 history but that
247 it diminished in later evolutionary periods. In addition, this ancient diversity could
248 still be underestimated due to under-sequencing of the geographic regions concerned
249 by this lineage. In other words, we would be dealing with a long history for this
250 lineage, which, because of a current disappearance context, would be hidden and
251 would pass "*under the radar*".

252 Regarding the genetic events having occurred along this diversification, we can
253 note that there are quite few IS*6110* insertions, RD, and missing genes. A notable
254 exception to the little role of IS*6110* is its insertion between spacers 30 and 35 in
255 CRISPR locus, which is a feature of the L5.1. This overall lack of L5 genomic
256 dynamism could contribute to its disappearance, likely because the bacillus is too
257 slow to adapt and because the long-lasting cross-talk between its host served the host
258 resistance. Hence, the phylogeny described in this paper is bound to evolve,
259 especially since L5 is severely undersampled for the time-being as compared to other
260 lineages.

261 We now want to remind of the provenance of samples and potential associated
262 biases. L5-caused tuberculosis represents an important part of all tuberculosis in some
263 important regions from Africa such as Nigeria, Benin, Cameroun, Ghana *i.e.* in
264 countries bordering the gulf of Guinea (Coscolla et al., 2021; Gehre et al., 2013;
265 Gygli et al., 2019; Osei-Wusu et al., 2021; Sanoussi et al., 2021; Silva et al., 2022).
266 But in Togo for example, the data of L5-L6 prevalence remains completely unknown.
267 In their 2013 paper describing the phylogeography of L5 in Benin, Nigeria, and Sierra

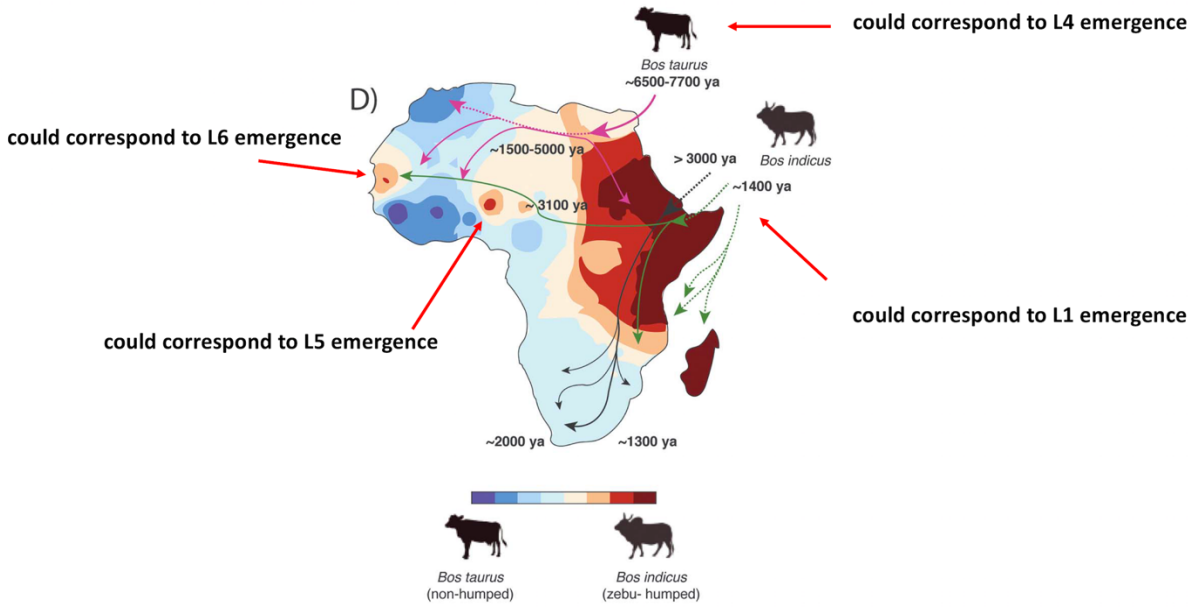
268 Leone and based on spoligotyping signatures, an attempt to describe 10 L5
269 sublineages, L5.1 to 5.10 was done, however without whole-genome sequencing
270 approach (Gehre et al., 2013). Actually, the landscape of CRISPR evolution in L5-
271 L6 is complex; from one hand, some sublineages can indeed be described, such as
272 the new L5.3 (new), that bears a yet unknown and rare specific 43 spacers spoligotype
273 (674077717777071). The former L5.3 (Coscolla et al., 2021) is now L5.2.3 and
274 should not considered as a subbranch of L5.2 (new). From another hand, the 43 spacer
275 signature is not always sufficient to assign definitively all other sublineages, although
276 many sublineages can be defined through specific spoligotyping signatures (see
277 **Supplementary Material 3**). We provide in this article a list of 45 SNPs that allows
278 to resolve the phylogenetic tree of L5 with 3 main branches (L5.1, L5.2, L5.3 (new)).
279 In some cases we link specific IS*6110* insertion/deletion events with some particular
280 sublineages emergence such as the loss of spacers 21-24 in relation to the loss of
281 IS*6110* in L5.1. Our result improves the former classification obtained by Ates et *al.*
282 and by Coscolla *et al.* (Ates et al., 2018; Coscolla et al., 2021). Still, a few strains
283 remain difficult to classify (**Figure 10**).

284 Our study still suffers from some limitations: when genomes are present in
285 databases, the geographical origin of the samples is often missing so that these data
286 are of poor interest for phylogeography. Moreover, data are either absent or scarce
287 for many countries of Africa, that still suffer from undescribed tuberculosis linked to
288 L5 or L6. A striking example is given by Nigeria, a 200 million people country that
289 has only two MTBC publicly available WGS in the NCBI genome database (search
290 done in July 2022). We also identified a trend for specific L5.1 and L5.2
291 phylogeographic distributions that remain to be confirmed by extensive recruitment
292 in central Africa. Whether this is caused by historical, environmental factors or
293 human genetics factors is however difficult to say. It is commonly accepted that
294 MTBC and its host developed under strong sympatry. This is however based on few
295 facts as research studies trying to link human genetics predisposition to bacterial
296 genetics remain rare (Gagneux et al., 2006).

297 Long-read sequencing on three geographically distant isolates of L5 suggested
298 that using H37Rv as a reference genome is limiting and showed that the pangenome
299 of L5 could be larger than expected previously (Sanoussi et al., 2021). Preliminary
300 results using *de novo* alignment using Roary indeed suggests that some L5 may

301 harbor specific genes that are not found in H37Rv (results not shown) (Page et al.,
302 2015). However the purpose of this study was not to focus on the pangenome, a
303 challenge that will require more L5 to be sequenced, but rather to focus on SNPs
304 analysis, CRISPR locus structure, classification and phylogeography of L5. The most
305 specific genes found to be absent in L5 and present in H37Rv are Rv1979c and
306 Rv2478c (results not shown).

307 Recently, an important paper by Kerner *et al*. studied ancient DNA (aDNA) of
308 Europeans to look for the time-distribution of an homozygous P1104A allele on the
309 tyrosine kinase TYK2 gene, an allele that has been found to be associated to an
310 increased susceptibility to tuberculosis in humans (Kerner et al., 2021). Using a
311 collection of more than 1000 ancient human DNA, they suggest that there was an
312 emergence of this allele around 30,000 years ago in west Eurasia, followed by a
313 negative selection that started around the bronze age in Europe (Kerner et al., 2021).
314 These data may also apply to Africa, but DNA from african human populations, *a*
315 *fortiori* aDNA are less studied for the time-being (Choudhury et al., 2020). A recent
316 study performed in Ghana suggested that there could be a strong anthropological
317 link between the patient group (in this case the Ewe) and the likelihood to have a
318 tuberculosis caused by *M. africanum* (Asante-Poku et al., 2015). Another recent
319 study suggests that patients infected with L5-L6 versus L4 possess distinct intestinal
320 microbiota (Namasivayam et al., 2020).



321

322 **Figure 11**: Hypothesis of emergence of L1-L5-L6 (adapted from
323 Campbell and Ranciaro 2021, reproduced with courtesy of the author)

An interesting but speculative issue concerns the historical timeframe and geographical area of emergence of L5 in a wider evolutionary context. Recently, a new phylogroup, named as MTBAP for "MTB-associated phylotype" was defined, as the "missing link" between a likely environmental ancestor and a specific host-adapted, professional pathogen mycobacteria that includes recently described new species such as *M. decipiens, M. shinjukuense, M. lacus and M. riyadhense* (Sapriel and Brosch, 2019). It is known for infectious diseases agents that, finding the center of diversity, or center of emergence, is not always simple, due to the importance of passed migratory events that may have blurred the original picture (Coscolla et al., 2021; Menardo et al., 2019). However slow-growing pathogens like the MTBC bacilli, living an intra-cellular life-style, that spread through aerosols and airways, that are pathogenic, are more likely to have left long-lasting scars and traces in animals and/or in human populations than rapid, extra-cellular, highly pathogenic pathogens, that are transmitted by an oro-fecal way, and that are more rapidly swept by new variants (Achtman, 2016). In this sense, L5 (and also L6) show an highly geographically-constrained patterns of interest: West of West-africa for L6 and Guinea gulf and central Africa for L5. Considering also the presence of L1 in East-Africa, likely to find its origin in South-Asia (Menardo et al., 2021), and the assumption that tuberculosis was a human-driven rather than a cattle-driven disease (Brosch et al., 2002), one must however consider with great attention the recent results obtained by Campbell and Ranciaro on lactase persistence in human beings in relation to the demographic history of pastoralists (Campbell and Ranciaro, 2021) (**Figure 11**). Whether lactase persistence in humans could be linked to L1-L4-L5-L6 MTBC lineages emergence in relation to cattle domestication remains an open question. Lactase persistence is linked both to animal herding and milk consumption. It has left traces of selection in humans and prevalence of the various adaptation alleles are also linked to the demographic histories of pastoralists population, in relation to climate change (Campbell and Ranciaro, 2021) unless gut microbial composition could compensate as in some pastorialist populations (Campbell and Ranciaro, 2021). This gut microbiota composition could also be associated to differential susceptibility to different lung infection agents (Namasivayam et al., 2020). Whether this is a cause or a consequence of infection remains to be studied,

357 however an indirect link may exist between a nomadic pastoralist and a certain type
358 of lung infections on one hand, and in another hand, a sedentary agricultural lifestyle
359 with other lung infection types. Related to this topic, a surprisingly congruent
360 geographical pattern is found between the presumed time-frame of emergence and
361 area of migration of the two main *Bos* species, *Bos taurus* and *Bos indicus*, and the
362 geographical distribution of L1, L4, L5, L6 in Africa (Campbell and Ranciaro, 2021).
363

364 We suggest that an MTBC lineage (L4) could have emerged in relation to a North-
365 african *Bos taurus* domestication, and hence, would be found only first in North of Africa and
366 by extension elsewhere in Africa, contemporary to the recent neolithic expansion (Bantu
367 agricultural expansion) (blue color in **Figure11**), an expansion that would have finally led to
368 a specialized human pathogen. Conversely, we suggest that the other L1-L5-L6 MTBC
369 lineages could have been related to *Bos indicus* domestication and introduction into Africa
370 via East-Africa and the arabian peninsula (red color in **Figure 11**). It is quite puzzling, as
371 shown in **Figure 11** that the two isolated red spots, indicator of *Bos indicus* introduction
372 events found in West Africa quite well correspond and can be superimposed to L5 and L6
373 geographic distribution (respectively Cameroun-Nigeria and Senegal-Gambia) and could
374 coincide with an early (3000 years ago) or late (>1400 years ago) arrival of *Bos indicus* from
375 the Arabic peninsula, first via Eastern Africa, and later arrival in central Africa and even later
376 in West Africa. Such a scenario also fits with the recent discovery that *M. bovis* also emerged
377 in Africa with nomad pastoralists having domesticated different cattle in the near-east or in
378 the Indus valley (Loiseau et al., 2020). Conversely the L4 sublineages could have spread from
379 West Africa to Central and South Africa following the Bantu speakers expansion (Patin et al.,
380 2017).

381 Hence, tuberculosis could either have initially been sitting in Africa since paleolithic
382 times in humns, or alternatively have been introduced into Africa many times through various
383 animal or human migrations routes in relation to the Neolithic transition, to human slave trade,
384 and possibly via South Asia and the Arabic peninsula (Comas et al., 2013; Menardo et al.,
385 2021), a complex history that remains to be better understood. L6 emergence is posterior to
386 L5 emergence and the fact that L5 and L6 are respectively more prone to be found in human
387 (L5) or in animal (L6), possibly argues of a potentially long duration of interactions of their
388 last common ancestor with animals and potential co-infection and complex cross-talk history
389 of tuberculosis in Africa (Perrin, 2015). Moreover in Africa, as already mentioned, the "*cattle*

390 *before crops*" model for food production is indeed very different from anywhere else in the
391 world (Manning, 2010). The neolithization process should also be understood differently in
392 Africa and in Europ, because of a very different history and a deep anthropological
393 significance (Manning, 2010).

394 To conclude, we believe that it is very likely that the early history of
395 tuberculosis caused by *M. africanum*, is intimately linked to local processes that are
396 linking demographical, anthropological, and ecological factors, including climate-
397 change, that shaped, throughout the millenias, today's Africa MTBC bacilli
398 population structure, with intricate relationships between human and animal life-
399 styles, from hunter-gatherer, pastoral, agro-pastoral, and more recently urban life-
400 styles. In Africa, an accurate reconstruction of the global remote and/or recent
401 tuberculosis history remains intimately linked to an expansion of numerous new local
402 epidemiological and evolutionary WGS studies. Nevertheless, in this article, we have
403 provided a more precise picture of some of the evolutionary steps of L5
404 diversification using for the largest L5 sample and a specific tool easing the
405 exploration of all genetic modifications along MTBC history, *TB-Annotator* (v.1,
406 15900 genomes and v.2 90000 genomes*)*. However, most of the functional
407 consequences of the identified genetic events characterized so far, still need to be
408 studied in more detail to grasp the virulence evolution of all these lineages.

409

410 ## List of Figures and tables

411 **Figure 1**: General algorithm of the TB-Annotator

412 **Figure 2**: General overview of the L5 phylogeny

413 **Figure 3**: Existing phylogeny (Coscolla et al., 2021) reference naming
414 and transient naming of new clades

415 **Figure 4**: New phylogeny with hierarchical level naming

416 **Figure 5**: 5.3 sublineage

417 **Figure 6**: 5.2.1 and 5.2.4 sublineages

418 **Figure 7**: L5.2.2 sublineage or L5.2 in (Coscolla et al.,
419 2021)

420 **Figure 8**: 5.2.3 sublineage

Figure 9: L5.1 and L5.2 geographical distribution as found in this study

Figure 10: A final phylogeny of Lineage 5 with SNP barcode proposal

Figure 11: Hypothesis of emergence of L1-L5-L6 and L4 (adapted from Campbell et al. 2021, reproduced with courtesy of the author)

Table 1: conversion table between the reference nomenclature and our new naming

Supplementary Material 1 Table 2: List of 45 SNPs, Lineage information at internal of leaf node, detail information on SNPs assessed and complementary informations (xls file)

Supplementary Material 2 A thorough classification of the L5 Sub-lineages (pdf. File, 10 pages).

Supplementary Material 3: Table 3: Accession List of genomes assessed, CRISPR-builder reconstructed spoligotype, countries of origin of the samples, miscellaneous information on CRISPR, comparison of classification results obtained between TB-Profiler lineage assignment, automatized spoligotype recorded in TB-Profiler, Final list of all spoligotypes (L5-spoligo-all), results of Ates et al. reanalyzed (xls. File, 47 pages)

Supplementary Material 4: Table 4: Accession List of genomes assessed, with all different L5 sublineages SNPs specificity search in TB-Annotator v1 (15900 genomes) (xls. File, 45 pages)

Supplementary Material 5: Two phylogenetical trees built using RaXML with or without PE-PPE gene variants on a subset of samples. (pdf. File, 3 pages).

Author statements

4.1 Author contributions

Conceptualization: CG, GS, GR, CS; data curation: CG, GS, CS, MRS, BMM; Formal analysis : CG,CS, MRS; Funding acquisition: CS, EC, CG; Investigation: CS, GR, CG, MRS, KL; Methodology: GS,GR,CG, MRS; Project Administration: TP,MRS,JD; Resources: EC;JD,CG,TP; Software: GS,KL, CG; Supervision:CS,CG, EC; Validation:

CS,KL,MRS; Visualization: CS,CG; Writing original. Manuscript; CG,CS,GR, MRS; Writing-review and editing: CG, CS, MRS, GR.

4.2 Conflicts of interest

The author(s) declare that there are no conflicts of interest

4.3 Funding information

This study was supported by PTDF (Petroleum Technology Development Trust) of Nigeria through a PhD grant attributed to MRS, by KNCV Nigeria (TP), by Campus-France through benchfees for MRS, by the UMR1137 IAME INSERM-Université Paris-City research Unit, and by the APHP (Assistance Publique des Hôpitaux de Paris) through the associated French NRL for TB. This project is part of MRS PhD (phylodynamics of MTBC with a specific focus on L5-L6 lineages).

4.4 Ethical approval

Public health action taken as a result of notification and surveillance is one of the tasks of the *Service de mycobactériologie spécialisée et de référence* of the Bichat Hospital in Paris, as a French TB NRL -associated laboratory, with identical missions as the one of NRL which provide a mandate and legislative basis to undertake necessary follow-up. Part of this follow-up is identification of epidemiological and molecular links between cases and research. This study is carried out under this framework, and as such explicit ethical approval was not mandatory at start working on archive material. However a prospective study, whose results are going to be reported elsewhere is currently under approval at the Nigerian Federal NRL ethical committee. Patient data are not traceable by any other third part than by the NRL. This research is done in the framework of the PhD thesis of MRS, grantee from the PTDF (Petroleum Technology Development Fund) of Nigeria.

4.5 Consent for publication

All authors read and approved the final version of this article.

4.6 Acknowledgements

This study is dedicated to Dr. Lovett Lawson, Nigeria STOP-TB officer, who passed away due to Covid19 in 2020. We are grateful to Barbara Molina-Moya, Jose Dominguez and Luis Cuevas for providing us with DNA archives of L5-L6 from the previous EDCTP

Collaborative study led by Pr. Luis Cuevas and Pr. Lovett Lawson. We are grateful to Dr. Jody Phelan, from the LSTM for providing us with an updated Spoligotyping list with associated SRA accession number. MRS is a PhD fellow from Nigeria supported by the Petroleum Technology Development Fund from Nigeria (PTDF) which is warmly acknowledged; Campus France is also acknowledged for his support and the IAME research Unit, UMR1137, and its director, Pr. Erik Denamur, for welcoming and supporting MRS.

References

Achtman, M., 2016. How old are bacterial pathogens? Proceedings. Biological sciences / The Royal Society 283.

Andrews, S., 2010. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online].

Asante-Poku, A., Otchere, I.D., Osei-Wusu, S., Sarpong, E., Baddoo, A., Forson, A., Laryea, C., Borrell, S., Bonsu, F., Hattendorf, J., Ahorlu, C., Koram, K.A., Gagneux, S., Yeboah-Manu, D., 2016. Molecular epidemiology of Mycobacterium africanum in Ghana. BMC Infect Dis 16, 385.

Asante-Poku, A., Yeboah-Manu, D., Otchere, I.D., Aboagye, S.Y., Stucki, D., Hattendorf, J., Borrell, S., Feldmann, J., Danso, E., Gagneux, S., 2015. Mycobacterium africanum is associated with patient ethnicity in Ghana. PLoS neglected tropical diseases 9, e3370.

Ates, L.S., Dippenaar, A., Sayes, F., Pawlik, A., Bouchier, C., Ma, L., Warren, R.M., Sougakoff, W., Majlessi, L., van Heijst, J.W.J., Brossier, F., Brosch, R., 2018. Unexpected Genomic and Phenotypic Diversity of Mycobacterium africanum Lineage 5 Affects Drug Resistance, Protein Secretion, and Immunogenicity. Genome biology and evolution 10, 1858-1874.

Brosch, R., Gordon, S.V., Marmiesse, M., Brodin, P., Buchrieser, C., Eiglmeier, K., Garnier, T., Gutierrez, C., Hewinson, G., Kremer, K., Parsons, L.M., Pym, A.S., Samper, S., van Soolingen, D., Cole, S.T., 2002. A new evolutionary scenario for the Mycobacterium tuberculosis complex. Proceedings of the National Academy of Sciences of the United States of America 99, 3684-3689.

Brudey, K., Gutierrez, M.C., Vincent, V., Parsons, L.M., Salfinger, M., Rastogi, N., Sola, C., 2004. Mycobacterium africanum Genotyping Using Novel Spacer Oligonucleotides in the Direct Repeat Locus. J Clin Microbiol 42, 5053-5057.

Campbell, M.C., Ranciaro, A., 2021. Human adaptation, demography and cattle domestication: an overview of the complexity of lactase persistence in Africa. Hum Mol Genet 30, R98-R109.

Castets, M., Boisvert, H., Grumbach, F., Brunel, M., Rist, N., 1968. Les bacilles tuberculeux de type africain: note préliminaire. Rev. Tuberc. Pneumol. 32, 179-184.

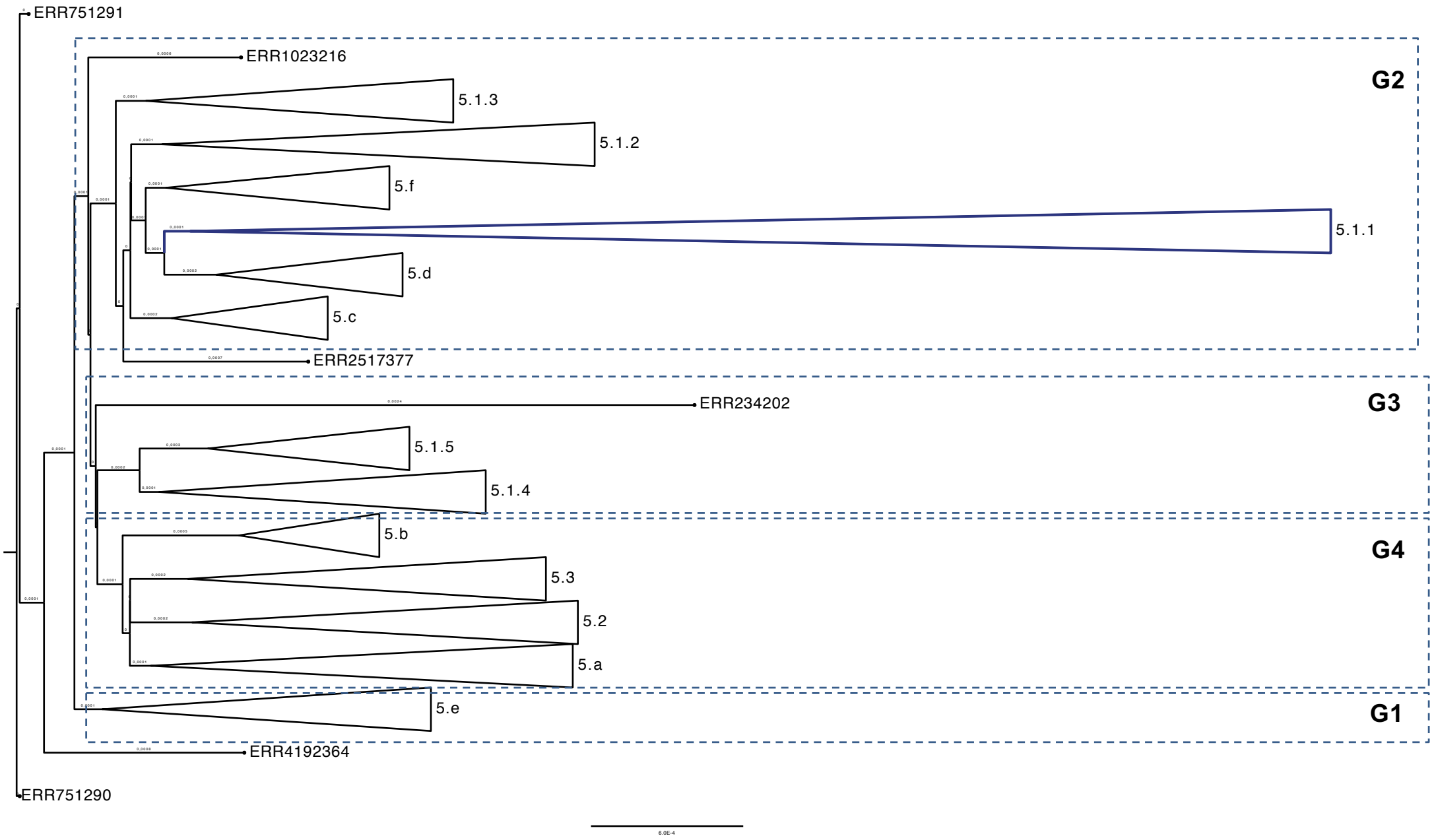
Choudhury, A., Aron, S., Botigue, L.R., Sengupta, D., Botha, G., Bensellak, T., Wells, G., Kumuthini, J., Shriner, D., Fakim, Y.J., Ghoorah, A.W., Dareng, E., Odia, T., Falola, O., Adebiyi, E., Hazelhurst, S., Mazandu, G., Nyangiri, O.A., Mbiyavanga, M., Benkahla, A., Kassim, S.K., Mulder, N., Adebamowo, S.N., Chimusa, E.R., Muzny, D., Metcalf, G., Gibbs, R.A., Trypano, G.E.N.R.G., Rotimi, C., Ramsay, M., Consortium, H.A., Adeyemo, A.A., Lombard, Z.,

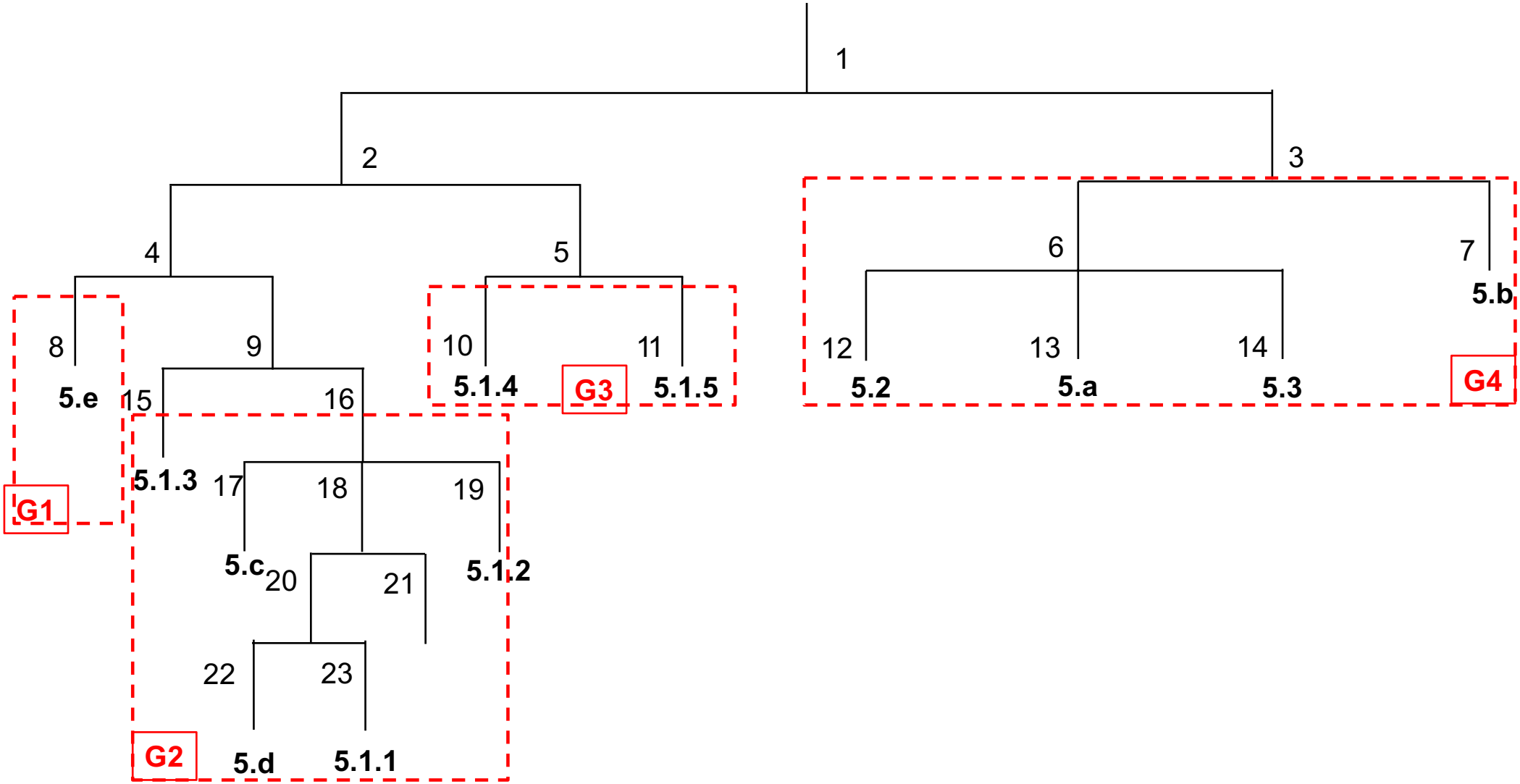
528 Hanchard, N.A., 2020. High-depth African genomes inform human migration and health.
529 Nature 586, 741-748.
530 Comas, I., Coscolla, M., Luo, T., Borrell, S., Holt, K.E., Kato-Maeda, M., Parkhill, J., Malla, B.,
531 Berg, S., Thwaites, G., Yeboah-Manu, D., Bothamley, G., Mei, J., Wei, L., Bentley, S., Harris,
532 S.R., Niemann, S., Diel, R., Aseffa, A., Gao, Q., Young, D., Gagneux, S., 2013. Out-of-Africa
533 migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans.
534 Nature genetics 45, 1176-1182.
535 Comas, I., Homolka, S., Niemann, S., Gagneux, S., 2009. Genotyping of genetically
536 monomorphic bacteria: DNA sequencing in Mycobacterium tuberculosis highlights the
537 limitations of current methodologies. PLoS One 4, e7815.
538 Coscolla, M., Gagneux, S., Menardo, F., Loiseau, C., Ruiz-Rodriguez, P., Borrell, S., Otchere,
539 I.D., Asante-Poku, A., Asare, P., Sanchez-Buso, L., Gehre, F., Sanoussi, C.N., Antonio, M.,
540 Affolabi, D., Fyfe, J., Beckert, P., Niemann, S., Alabi, A.S., Grobusch, M.P., Kobbe, R., Parkhill,
541 J., Beisel, C., Fenner, L., Bottger, E.C., Meehan, C.J., Harris, S.R., de Jong, B.C., Yeboah-Manu,
542 D., Brites, D., 2021. Phylogenomics of Mycobacterium africanum reveals a new lineage and a
543 complex evolutionary history. Microb Genom 7.
544 Couvin, D., David, A., Zozio, T., Rastogi, N., 2019. Macro-geographical specificities of the
545 prevailing tuberculosis epidemic as seen through SITVIT2, an updated version of the
546 Mycobacterium tuberculosis genotyping database. Infect Genet Evol 72, 31-43.
547 David, H.L., Jahan, M.T., Jumin, A., Grandry, J., Lehmann, E.H., 1978. Numerical taxonomy of
548 *Mycobacterium africanum*. Int. J. System. Bacteriol. 28, 467-472.
549 de Jong, B.C., Antonio, M., Awine, T., Ogungbemi, K., de Jong, Y.P., Gagneux, S., DeRiemer,
550 K., Zozio, T., Rastogi, N., Borgdorff, M., Hill, P.C., Adegbola, R.A., 2009. Use of spoligotyping
551 and large sequence polymorphisms to study the population structure of the Mycobacterium
552 tuberculosis complex in a cohort study of consecutive smear-positive tuberculosis cases in
553 The Gambia. J Clin Microbiol 47, 994-1001.
554 Gagneux, S., DeRiemer, K., Van, T., Kato-Maeda, M., de Jong, B.C., Narayanan, S., Nicol, M.,
555 Niemann, S., Kremer, K., Gutierrez, M.C., Hilty, M., Hopewell, P.C., Small, P.M., 2006.
556 Variable host-pathogen compatibility in Mycobacterium tuberculosis. Proceedings of the
557 National Academy of Sciences of the United States of America 103, 2869-2873.
558 Gehre, F., Antonio, M., Faihun, F., Odoun, M., Uwizeye, C., de Rijk, P., de Jong, B.C., Affolabi,
559 D., 2013. The first phylogeographic population structure and analysis of transmission
560 dynamics of M. africanum West African 1--combining molecular data from Benin, Nigeria
561 and Sierra Leone. PLoS One 8, e77000.
562 Guyeux, C., Senelle, G., Refrégier, G., Bretelle-Establet, F., Cambau, E., Sola, C., 2022.
563 Connection between two historical tuberculosis outbreak sites in Japan, Honshu, by a new
564 ancestral Mycobacterium tuberculosis L2 sublineage Epidemiology and infection 150, 1-10.
565 Guyeux, C., Sola, C., Nous, C., Refregier, G., 2021. CRISPRbuilder-TB: "CRISPR-builder for
566 tuberculosis". Exhaustive reconstruction of the CRISPR locus in mycobacterium tuberculosis
567 complex using SRA. PLoS computational biology 17, e1008500.
568 Gygli, S.M., Keller, P.M., Ballif, M., Blochliger, N., Homke, R., Reinhard, M., Loiseau, C.,
569 Ritter, C., Sander, P., Borrell, S., Collantes Loo, J., Avihingsanon, A., Gnokoro, J., Yotebieng,
570 M., Egger, M., Gagneux, S., Bottger, E.C., 2019. Whole-Genome Sequencing for Drug
571 Resistance Profile Prediction in Mycobacterium tuberculosis. Antimicrob Agents Chemother
572 63.
573 Holmes, J.B., Moyer, E., Phan, L., Maglott, D., Kattman, B., 2020. SPDI: data model for
574 variants and applications at NCBI. Bioinformatics 36, 1902-1907.

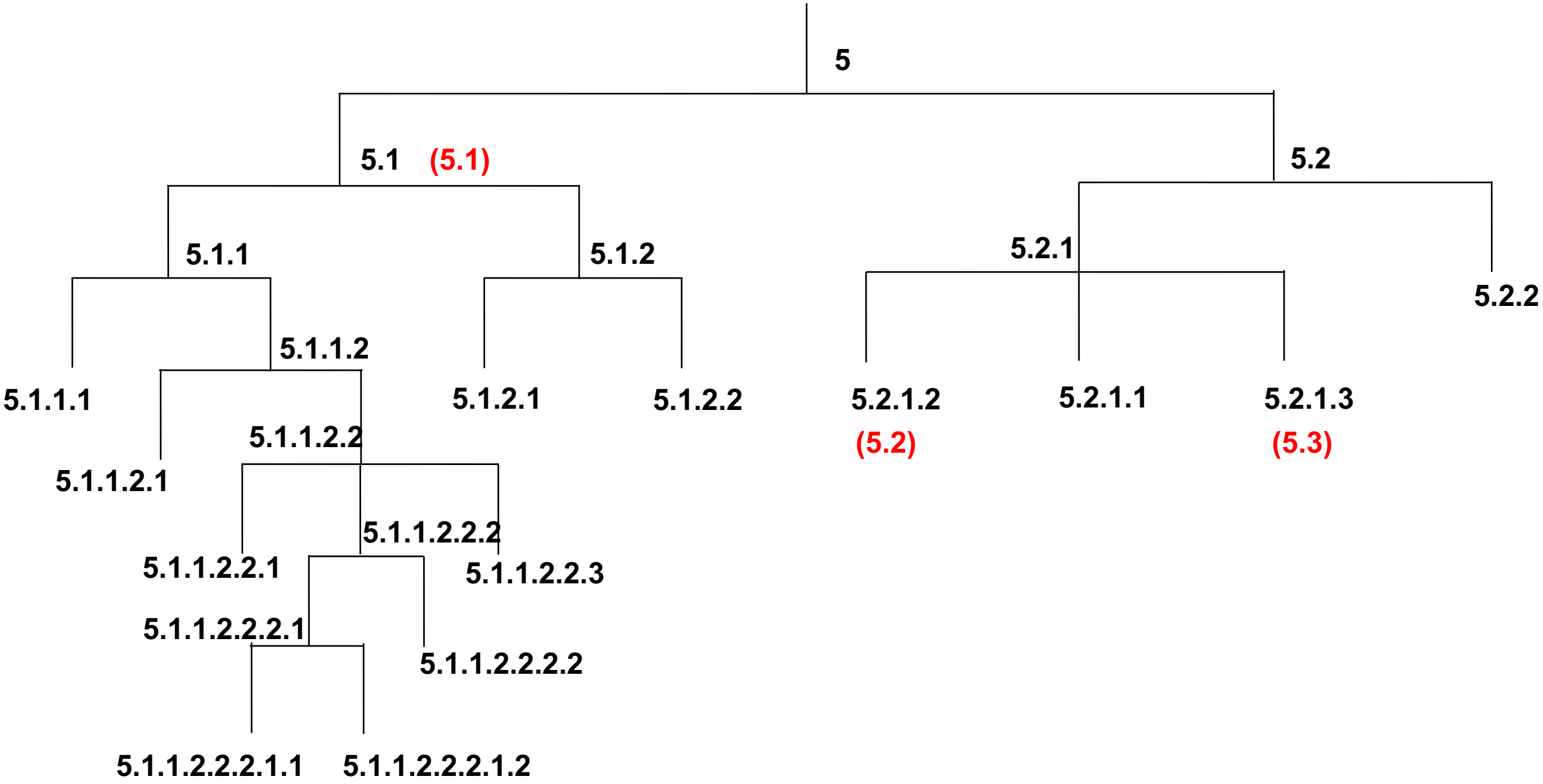
575 Kamerbeek, J., Schouls, L., Kolk, A., van Agterveld, M., van Soolingen, D., Kuijper, S.,
576 Bunschoten, A., Molhuizen, H., Shaw, R., Goyal, M., van Embden, J., 1997. Simultaneous
577 detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and
578 epidemiology. J Clin Microbiol 35, 907-914.
579 Kerner, G., Laval, G., Patin, E., Boisson-Dupuis, S., Abel, L., Casanova, J.L., Quintana-Murci, L.,
580 2021. Human ancient DNA analyses reveal the high burden of tuberculosis in Europeans
581 over the last 2,000 years. American journal of human genetics 108, 517-524.
582 Kuper, R., Kropelin, S., 2006. Climate-controlled Holocene occupation in the Sahara: motor
583 of Africa's evolution. Science 313, 803-807.
584 Lawson, L., Zhang, J., Gomgnimbou, M.K., Abdurrahman, S.T., Le Moullec, S., Mohamed, F.,
585 Uzoewulu, G.N., Sogaolu, O.M., Goh, K.S., Emenyonu, N., Refregier, G., Cuevas, L.E., Sola, C.,
586 2012. A molecular epidemiological and genetic diversity study of tuberculosis in ibadan,
587 nnewi and abuja, Nigeria. PLoS One 7, e38409.
588 Loiseau, C., Menardo, F., Aseffa, A., Hailu, E., Gumi, B., Ameni, G., Berg, S., Rigouts, L.,
589 Robbe-Austerman, S., Zinsstag, J., Gagneux, S., Brites, D., 2020. An African origin for
590 Mycobacterium bovis. Evolution, medicine, and public health 2020, 49-59.
591 Malm, S., Linguissi, L.S., Tekwu, E.M., Vouvoungui, J.C., Kohl, T.A., Beckert, P., Sidibe, A.,
592 Rusch-Gerdes, S., Madzou-Laboum, I.K., Kwedi, S., Penlap Beng, V., Frank, M., Ntoumi, F.,
593 Niemann, S., 2017. New Mycobacterium tuberculosis Complex Sublineage, Brazzaville,
594 Congo. Emerg Infect Dis 23, 423-429.
595 Manning, K., 2010. A developmental history of West African agriculture, in: Allsworth-Jones,
596 P. (Ed.), West African Archaeology New developments, new perspectives. Hadrian Books,
597 Ltd, Archaeopress, Publishers of British Archaeological Reports, BAR International Series
598 2164, Oxford, pp. 43-52.
599 Menardo, F., Duchene, S., Brites, D., Gagneux, S., 2019. The molecular clock of
600 Mycobacterium tuberculosis. PLoS pathogens 15, e1008067.
601 Menardo, F., Rutaihwa, L., Zwyer, M., Borrell, S., Comas, I., ConceiÁ„o, E., Coscolla, M., Cox,
602 H., Joloba, M., Dou, H., Feldmann, J., Fenner, L., Fyfe, J., Gao, Q., GarcÌa de Viedma, D.,
603 Garcia-Basteiro, A., Gygli, S., Hella, J., Hiza, H., Jugheli, L., Kamwela, L., Kato-Maeda, M., Liu,
604 Q., Ley, S., Loiseau, C., Mahasirimongkol, S., Malla, B., Palittapongarnpim, P.,
605 Rakotosamimanana, N., Rasolofo, V., Reinhard, M., Reither, K., Sasamalo, M., Silva Duarte,
606 R., Sola, C., Suffys, P., Batista Lima, K., Yeboah-Manu, D., Beisel, C., Brites, D., Gagneux, S.,
607 2021. Local adaptation in populations of Mycobacterium tuberculosis endemic to the Indian
608 Ocean Rim [version 2; peer review: 2 approved]. F1000Research 10.
609 Molina-Moya, B., Abdurrahman, S.T., Madukaji, L.I., Gomgnimbou, M.K., Spinasse, L.,
610 Gomes-Fernandes, M., Gomes, H.M., Kacimi, S., Dacombe, R., Bimba, J.S., Lawson, L., Sola,
611 C., Cuevas, L.E., Dominguez, J., 2018a. Genetic characterization of Mycobacterium
612 tuberculosis complex isolates circulating in Abuja, Nigeria. Infect Drug Resist 11, 1617-1625.
613 Molina-Moya, B., Gomgnimbou, M.K., Spinasse, L., Obasanya, J., Oladimeji, O., Dacombe, R.,
614 Edwards, T., Daragon, X.O., Lawson, L., Abdurrahman, S.T., Cuevas, L.E., Dominguez, J., Sola,
615 C., 2018b. Mycobacterium tuberculosis complex genotypes circulating in Nigeria based on
616 spoligotyping obtained from Ziehl-Neelsen stained slides extracted DNA. PLoS neglected
617 tropical diseases 12, e0006242.
618 Mostowy, S., Onipede, A., Gagneux, S., Niemann, S., Kremer, K., Desmond, E.P., Kato-
619 Maeda, M., Behr, M., 2004. Genomic Analysis Distinguishes Mycobacterium africanum. J
620 Clin Microbiol 42, 3594-3599.

621 Namasivayam, S., Diarra, B., Diabate, S., Sarro, Y.D.S., Kone, A., Kone, B., Tolofoudie, M.,
622 Baya, B., Diakite, M.T., Kodio, O., Cohen, K., Holl, J., Achenbach, C.J., Chatterjee, S., Murphy,
623 R.L., Bishai, W., Diallo, S., Sher, A., Maiga, M., 2020. Patients infected with Mycobacterium
624 africanum versus Mycobacterium tuberculosis possess distinct intestinal microbiota. PLoS
625 neglected tropical diseases 14, e0008230.
626 Napier, G., Campino, S., Merid, Y., Abebe, M., Woldeamanuel, Y., Aseffa, A., Hibberd, M.L.,
627 Phelan, J., Clark, T.G., 2020. Robust barcoding and identification of Mycobacterium
628 tuberculosis lineages for epidemiological and clinical studies. Genome medicine 12, 114.
629 Niobe-Eyangoh, S.N., Kuaban, C., Sorlin, P., Cunin, P., Thonnon, J., Sola, C., Rastogi, N.,
630 Vincent, V., Gutierrez, M.C., 2003. Genetic Biodiversity of Mycobacterium tuberculosis
631 Complex Strains from Patients with Pulmonary Tuberculosis in Cameroon. Journal of Clinical
632 Microbiology 41, 2547-2553.
633 Osei-Wusu, S., Otchere, I.D., Morgan, P., Musah, A.B., Siam, I.M., Asandem, D., Afum, T.,
634 Asare, P., Asante-Poku, A., Kusi, K.A., Gagneux, S., Yeboah-Manu, D., 2021. Genotypic and
635 phenotypic diversity of Mycobacterium tuberculosis complex genotypes prevalent in West
636 Africa. PLoS One 16, e0255433.
637 Otchere, I.D., Coscollá, M., Sánchez-Busó, L., Asante-Poku, A., Brites, D., Loiseau, C.,
638 Meehan, C., Osei-Wusu, S., Forson, A., Laryea, C., Yahayah, A.I., Baddoo, A., Ansa, G.A.,
639 Aboagye, S.Y., Asare, P., Borrell, S., Gehre, F., Beckert, P., Kohl, T.A., N'Dira, S., Beisel, C.,
640 Antonio, M., Niemann, S., de Jong, B.C., Parkhill, J., Harris, S.R., Gagneux, S., Yeboah-Manu,
641 D., 2018. Comparative genomics of Mycobacterium africanum Lineage 5 and Lineage 6 from
642 Ghana suggests distinct ecological niches. Scientific reports 8, 11269.
643 Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T., Fookes, M.,
644 Falush, D., Keane, J.A., Parkhill, J., 2015. Roary: rapid large-scale prokaryote pan genome
645 analysis. Bioinformatics 31, 3691-3693.
646 Patin, E., Lopez, M., Grollemund, R., Verdu, P., Harmant, C., Quach, H., Laval, G., Perry, G.H.,
647 Barreiro, L.B., Froment, A., Heyer, E., Massougbodji, A., Fortes-Lima, C., Migot-Nabias, F.,
648 Bellis, G., Dugoujon, J.M., Pereira, J.B., Fernandes, V., Pereira, L., Van der Veen, L.,
649 Mouguiama-Daouda, P., Bustamante, C.D., Hombert, J.M., Quintana-Murci, L., 2017.
650 Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North
651 America. Science 356, 543-546.
652 Perrin, P., 2015. Human and tuberculosis co-evolution: An integrative view. Tuberculosis
653 (Edinb).
654 Phillipson, D.W., 1975. The chronology of the Iron Age in Bantu Africa. J. Afr. Hist. 16, 321-
655 342.
656 Refrégier, G., Sola, C., Guyeux, C., 2020. Unexpected diversity of CRISPR unveils some
657 evolutionary patterns of repeated sequences in Mycobacterium tuberculosis. BMC genomics
658 21, 841.
659 Sanoussi, C.N., Coscolla, M., Ofori-Anyinam, B., Otchere, I.D., Antonio, M., Niemann, S.,
660 Parkhill, J., Harris, S., Yeboah-Manu, D., Gagneux, S., Rigouts, L., Affolabi, D., de Jong, B.C.,
661 Meehan, C.J., 2021. Mycobacterium tuberculosis complex lineage 5 exhibits high levels of
662 within-lineage genomic diversity and differing gene content compared to the type strain
663 H37Rv. Microb Genom 7.
664 Sapriel, G., Brosch, R., 2019. Shared Pathogenomic Patterns Characterize a New Phylotype,
665 Revealing Transition toward Host-Adaptation Long before Speciation of Mycobacterium
666 tuberculosis. Genome biology and evolution 11, 2420-2438.

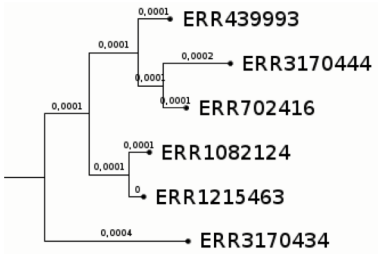
667 Silva, M.L., Ca, B., Osorio, N.S., Rodrigues, P.N.S., Maceiras, A.R., Saraiva, M., 2022.
668 Tuberculosis caused by Mycobacterium africanum: Knowns and unknowns. PLoS pathogens
669 18, e1010490.
670 Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
671 large phylogenies. Bioinformatics 30, 1312-1313.
672 van Embden, J.D.A., van Gorkom, T., Kremer, K., Jansen, R., van der Zeijst, B.A.M., Schouls,
673 L.M., 2000. Genetic variation and evolutionary origin of the Direct repeat locus of
674 *Mycobacterium tuberculosis* complex bacteria. J. Bacteriol. 182, 2393-2401.
675 Wamala, D., Okee, M., Kigozi, E., Couvin, D., Rastogi, N., Joloba, M., Kallenius, G., 2015.
676 Predominance of Uganda genotype of Mycobacterium tuberculosis isolated from Ugandan
677 patients with tuberculous lymphadenitis. BMC Res Notes 8, 398.
678 Yimer, S.A., Kalayou, S., Homberset, H., Birhanu, A.G., Riaz, T., Zegeye, E.D., Lutter, T.,
679 Abebe, M., Holm-Hansen, C., Aseffa, A., Tonjum, T., 2020. Lineage-Specific Proteomic
680 Signatures in the Mycobacterium tuberculosis Complex Reveal Differential Abundance of
681 Proteins Involved in Virulence, DNA Repair, CRISPR-Cas, Bioenergetics and Lipid Metabolism.
682 Frontiers in microbiology 11, 550760.
683

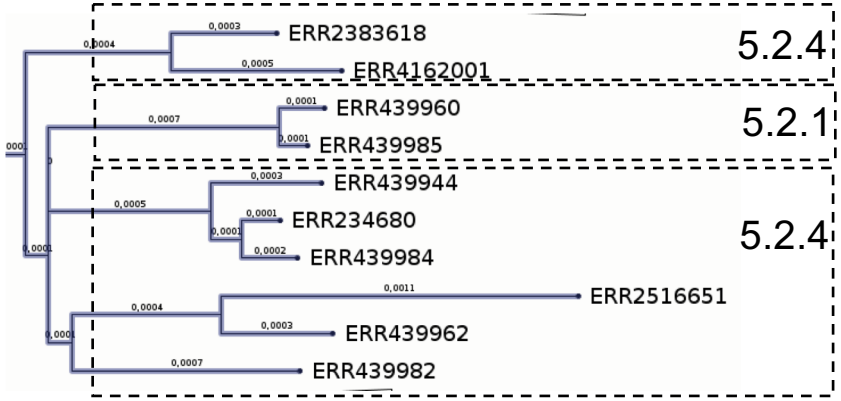

ERR751291
ERR1023216
G2
5.1.3
5.1.2
5.f
5.1.1
5.d
5.c
ERR2517377
ERR234202
G3
5.1.5
5.1.4
5.b
G4
5.3
5.2
5.a
5.e
G1
ERR4192364
ERR751290
6.0E-4

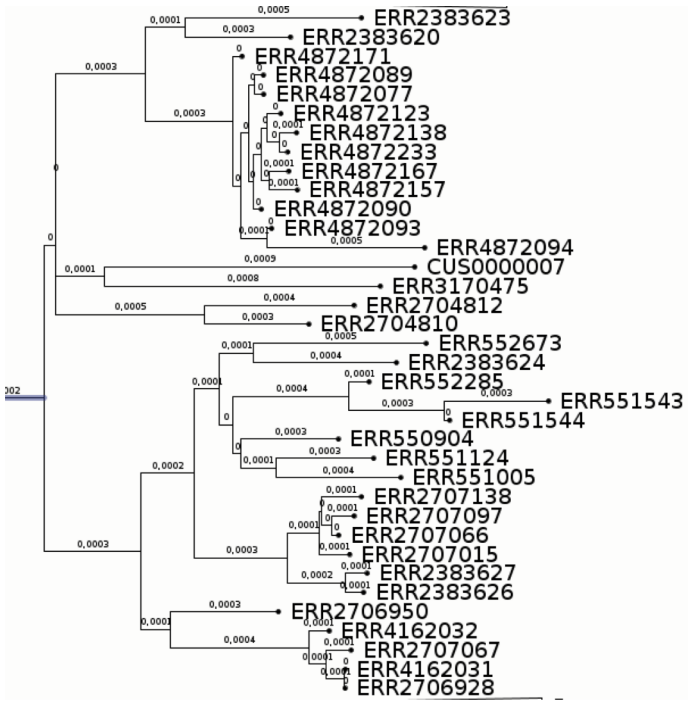

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
5.e
5.1.4
5.1.5
5.b
5.2
5.a
5.3
5.1.3
5.c
5.1.2
5.d
5.1.1
G1
G2
G3
G4

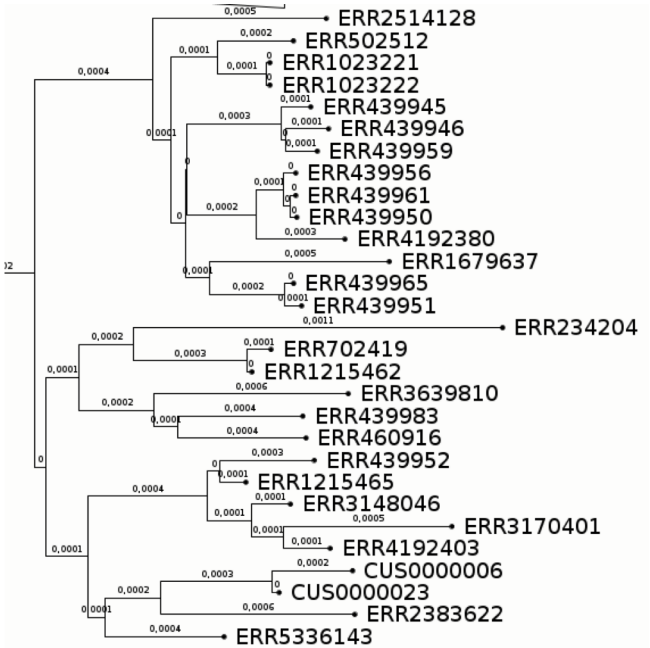


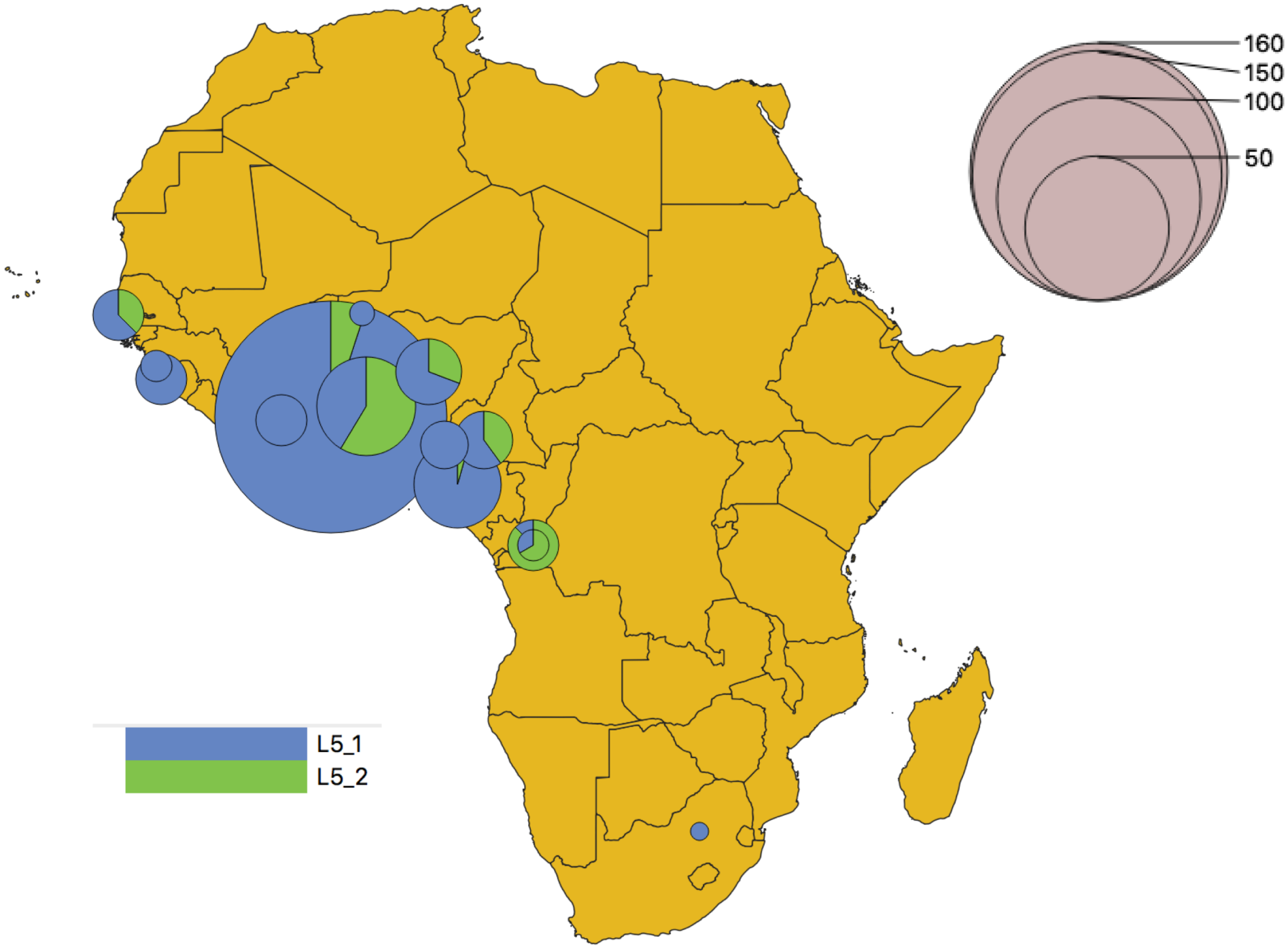
in bracket : former naming


0,0001
ERR439993
0,0001
0,0002
ERR3170444
0,0001
0,0001
ERR702416
0,0001
0,0001
ERR1082124
0,0001
0
ERR1215463
0,0004
ERR3170434

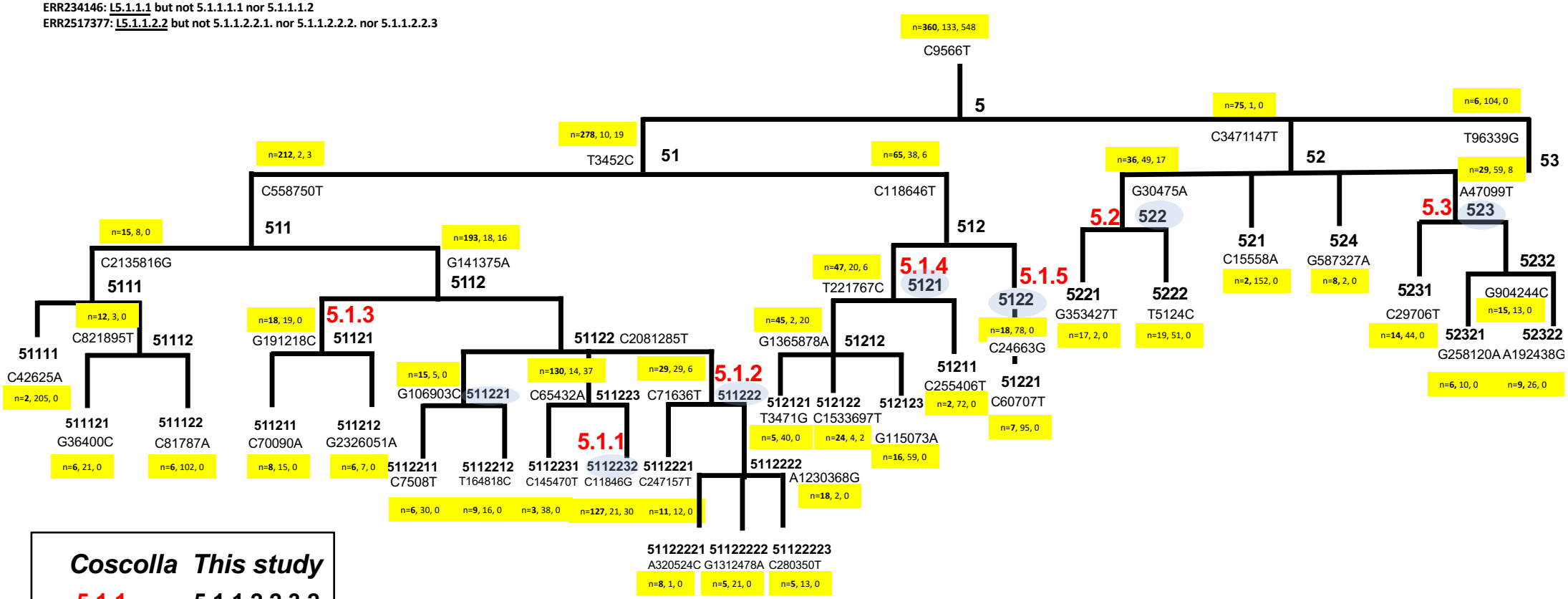

ERR2383618
ERR4162001
5.2.4
ERR439960
ERR439985
5.2.1
ERR439944
ERR234680
ERR439984
5.2.4
ERR2516651
ERR439962
ERR439982
0,0003
0,0004
0,0005
0,0001
0,0001
0,0007
0,0001
0,0003
0,0005
0,0001
0,0001
0,0002
0,0011
0,0004
0,0003
0,0007


ERR2383623
ERR2383620
ERR4872171
ERR4872089
ERR4872077
ERR4872123
ERR4872138
ERR4872233
ERR4872167
ERR4872157
ERR4872090
ERR4872093
ERR4872094
CUS0000007
ERR3170475
ERR2704812
ERR2704810
ERR552673
ERR2383624
ERR552285
ERR551543
ERR551544
ERR550904
ERR551124
ERR551005
ERR2707138
ERR2707097
ERR2707066
ERR2707015
ERR2383627
ERR2383626
ERR2706950
ERR4162032
ERR2707067
ERR4162031
ERR2706928


ERR2514128
ERR502512
ERR1023221
ERR1023222
ERR439945
ERR439946
ERR439959
ERR439956
ERR439961
ERR439950
ERR4192380
ERR1679637
ERR439965
ERR439951
ERR234204
ERR702419
ERR1215462
ERR3639810
ERR439983
ERR460916
ERR439952
ERR1215465
ERR3148046
ERR3170401
ERR4192403
CUS0000006
CUS0000023
ERR2383622
ERR5336143

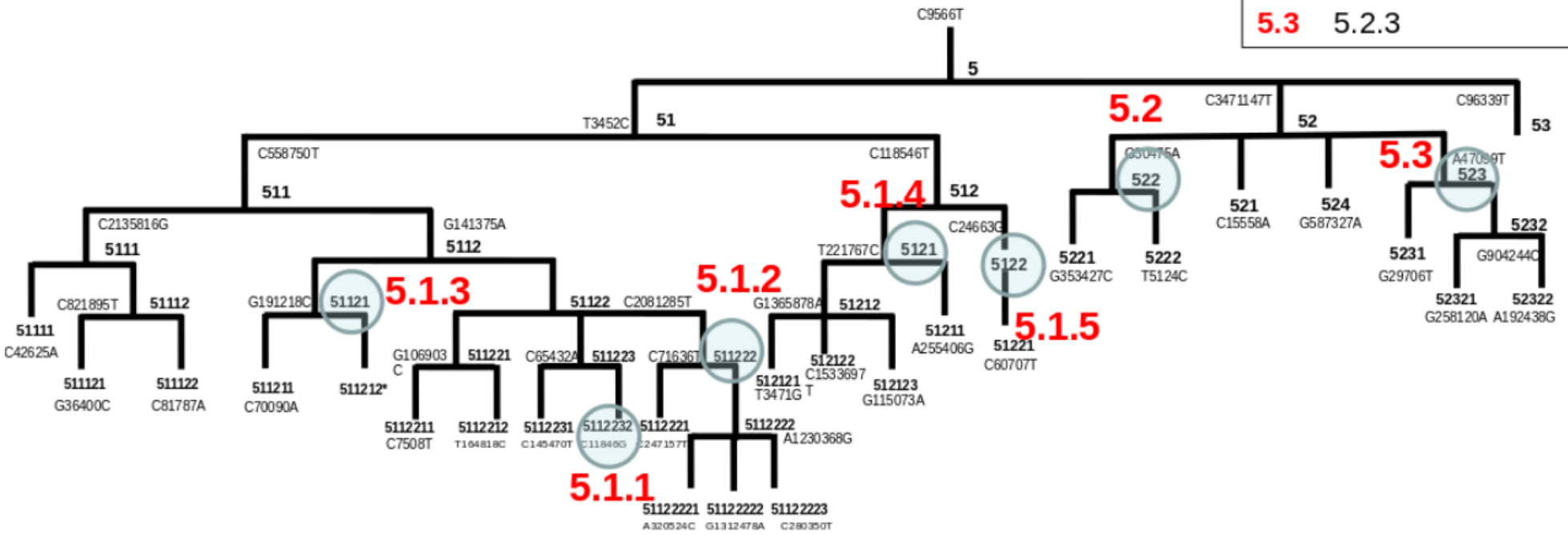

160
150
100
50
L5_1
L5_2

ERR4872200 (CMR) : L5 but not 5.1, not 5.2, not 5.3
ERR234202 : L5.1 but not 5.1.1 and not 5.1.2
ERR1023216, ERR4192364, ERR751290, ERR751291: L5.1.1 but not 5.1.1.1 nor 5.1.1.2
ERR234146: L5.1.1.1 but not 5.1.1.1.1 nor 5.1.1.1.2
ERR2517377: L5.1.1.2.2 but not 5.1.1.2.2.1. nor 5.1.1.2.2.2. nor 5.1.1.2.2.3



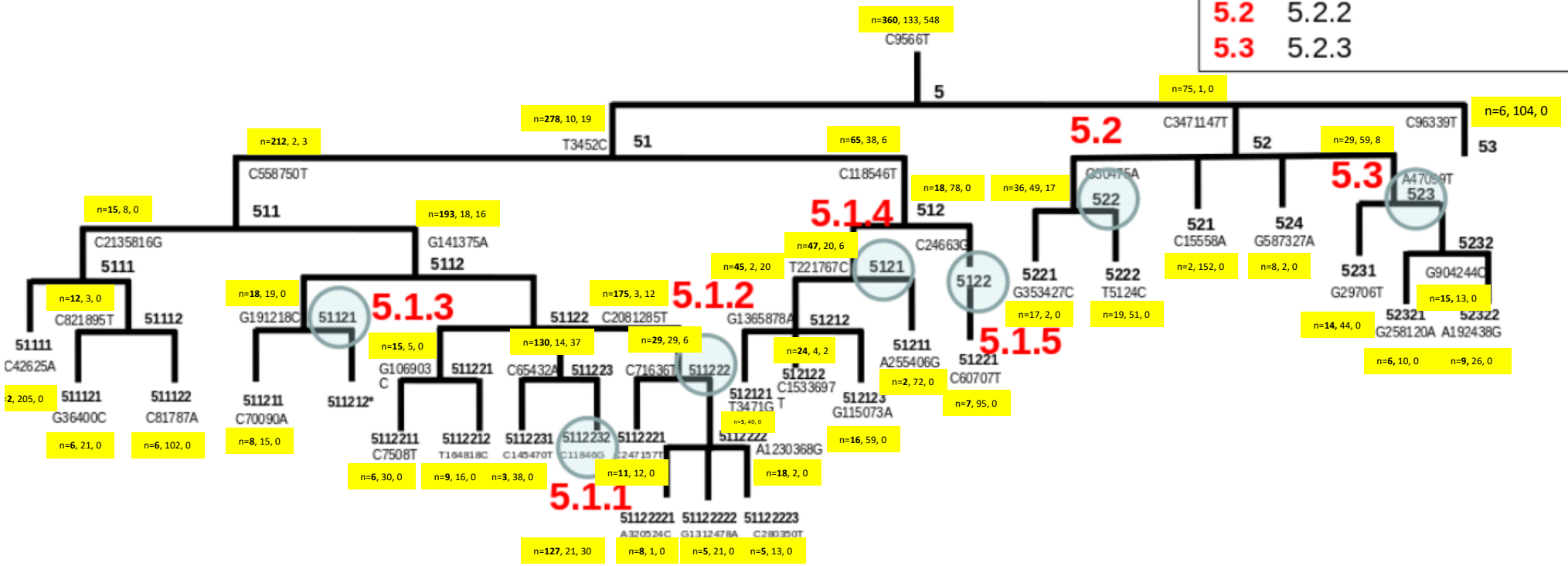
Coscolla	*This study*
5.1.1	5.1.1.2.2.3.2
5.1.2	5.1.1.2.2.2
5.1.3	5.1.1.2.1
5.1.4	5.1.2.1
5.1.5	5.1.2.2
5.2	5.2.2
5.3	5.2.3

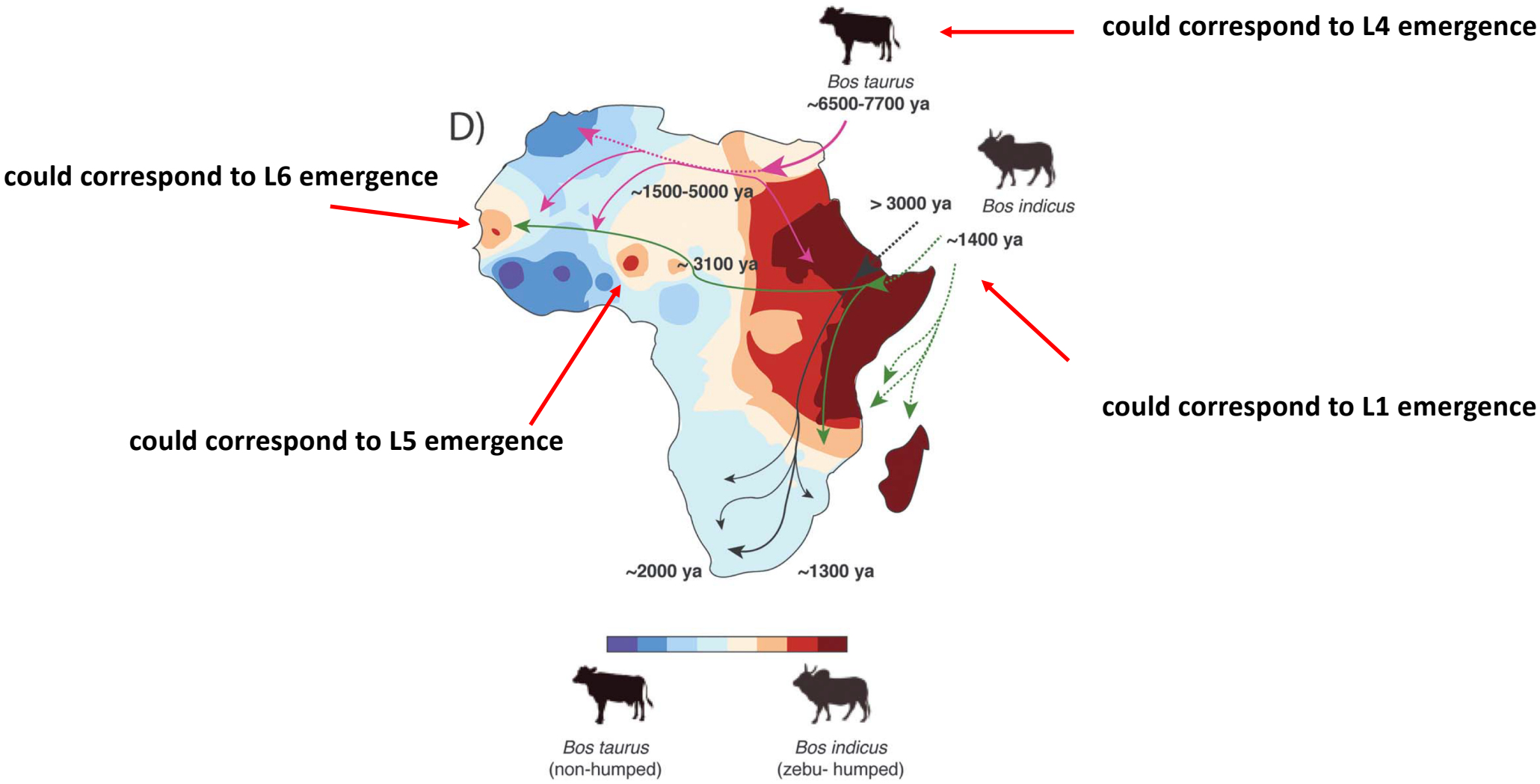
Coscolla	Our Denomination
5.1.1	5.1.1.2.2.3.2
5.1.2	5.1.1.2.2.2
5.1.3	5.1.1.2.1
5.1.4	5.1.2.1
5.1.5	5.1.2.2
5.2	5.2.2
5.3	5.2.3

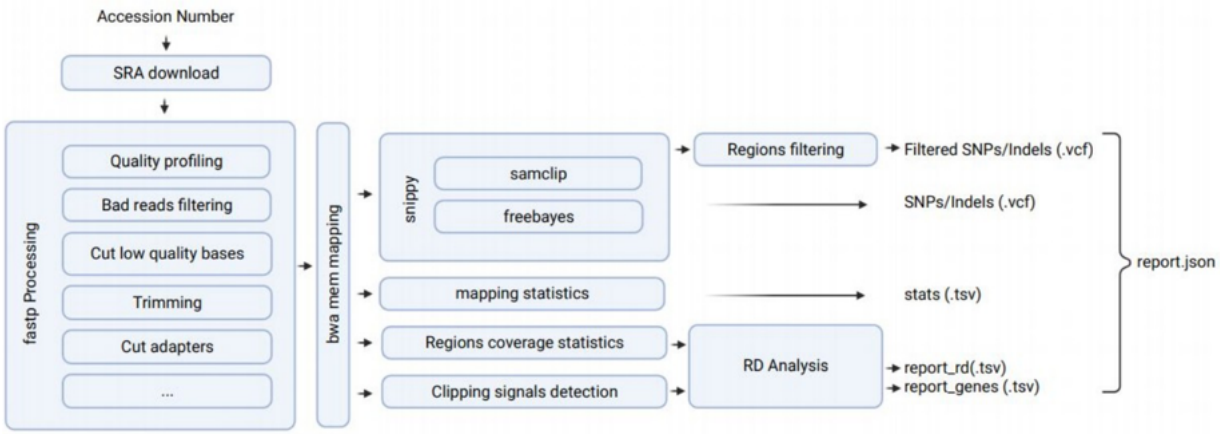


ERR4872200: L5 but not 5.1, not 5.2, not 5.3
ERR234202 : L5.1 but not 5.1.1 and not 5.1.2
ERR1023216, ERR4192364, ERR751290, ERR751291: L5.1.1 but not 5.1.1.1 nor 5.1.12
ERR234146: L5.1.1.1 but not 5.1.1.1.1 nor 5.1.1.1.2
ERR2517377: L5.1.1.2.2 but not 5.1.1.2.2.1. nor 5.1.1.2.2.2. nor 5.1.1.2.2.3

Coscolla	Our Denomination
5.1.1	5.1.1.2.2.3.2
5.1.2	5.1.1.2.2.2
5.1.3	5.1.1.2.1
5.1.4	5.1.2.1
5.1.5	5.1.2.2
5.2	5.2.2
5.3	5.2.3




could correspond to L4 emergence
Bos taurus
~6500-7700 ya
D)
could correspond to L6 emergence
~1500-5000 ya
> 3000 ya
Bos indicus
~1400 ya
~3100 ya
could correspond to L1 emergence
could correspond to L5 emergence
~2000 ya
~1300 ya
Bos taurus
(non-humped)
Bos indicus
(zebu- humped)


Accession Number
SRA download
fastp Processing
Quality profiling
Bad reads filtering
Cut low quality bases
Trimming
Cut adapters
...
bwa mem mapping
snippy
samclip
freebayes
mapping statistics
Regions coverage statistics
Clipping signals detection
Regions filtering
RD Analysis
Filtered SNPs/Indels (.vcf)
SNPs/Indels (.vcf)
stats (.tsv)
report_rd(.tsv)
report_genes (.tsv)
report.json