IV ETUDE 2 : DISTRIBUTION, DIVERSITE ET TRANSMISSION DE LA SOUCHE *M. tuberculosis* ENDEMIQUE MALGACHE : LE SIT109 (Article 2)

IV.1 PRESENTATION DE L'ETUDE

La phylogéographie des souches *M. tuberculosis* a démontré l'association des lignées de souches avec leur lieu de dispersion ou d'endémicité. De précédentes études ont montré l'existence d'une souche endémique et prédominante à Madagascar : Le SIT109. Le SIT109 appartient à la lignée EAI et à la sous-lignée des EAI8. Jusqu'à présent, aucune étude n'a jamais été faite sur ces souches. Cette étude constitue la première étude analysant le niveau de diversité des souches sous-jacentes à ce spoligotype endémique. L'objectif de l'étude étant d'étudier la diversité des SIT109 disponibles (n=156) au laboratoire des Mycobactéries de l'IPM par des méthodes de typage plus discriminantes du BK (le spoligotypage avec 68 espaceurs et les MIRU-VNTR). L'objectif secondaire étant de proposer une sélection minimale de locus VNTR pouvant typer les SIT109 avec un maximum de niveau de discrimination.

IV.2 ARTICLE 2

Genetic Diversity and Hypothetical Origin of the L1/SIT 109 Malagasy *Mycobacterium tuberculosis* Clinical Isolates

Noël H. Ratovonirina [*1, 2], Guislaine Refregier[2], Voahangy Rasolofo-Razanamparany[1], Christophe Sola*[2]

[1] Institut Pasteur de Madagascar, Unité des Mycobactéries

[2] Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91198, Gif-sur-Yvette cedex, France.

Corresponding authors:

***e-mail: ,** harijaona@pasteur.mg, christophe.sola@i2bc.paris-saclay.fr

Short title: ***M. tuberculosis*** **L1/SIT109 in Madagascar**

ABSTRACT (249 words)

Previous studies demonstrated the presence of an endemic Lineage 1 *Mycobacterium tuberculosis* (MTB) strain circulating predominantly in Madagascar, the spoligo-international-type 109 (L1/SIT109). Until now, very little knowledge about the L1/SIT109 genetic diversity and the origin of this genotype was available. The aim of this study was to evaluate the genetic diversity level of the L1/SIT109 sublineage using more discriminant *M. tuberculosis* genotyping methods, and try to hypothesize on its origin.

To achieve this goal, a first sample of 53 L1/SIT109 isolates available at the Institut Pasteur of Madagascar were typed by the extended 68 spacers spoligotyping method using a high throughput method and using complementary 24 MIRU-VNTR typing. In parallel, a selection of the 10 most discriminant MIRU-VNTR loci among the 24 was chosen to assess a MIRU-VNTR genotyping method to evaluate the genetic diversity on a larger collection of clinical isolates (n=103). Results showed that spoligotypes with 68 spacers did not discriminate L1/SIT109 clinical isolates (HGDI=0.097). Only 4 spoligotypes were discriminated with 1 single cluster of 50 isolates and 3 unique spoligotypes, whereas 24 MIRU-VNTR showed a larger genetic diversity of the L1/SIT109 (41 profiles of which 36 unique patterns and 5 clusters of 2 to 7 isolates) with an HGDI of 0.978. The result of the 24 MIRU-VNTR typing showed 9 invariable loci (MIRU03, MIRU20, MIRU24, MIRU26, Mtub04, Mtub29, Mtu30, Mtub34 and Qub4156). The selection of the 10 most discriminant loci (ETRA, ETRB, ETRD, MIRU10, ETRE, MIRU40, Mtub21, Mtub39, Qub11b and Qub26; HGDI values: 0.142 to 0.610) was used to subtype all of the L1/SIT109 isolates (n=156) with a similar discrimination level (HGDI=0.981). Finally we genotyped a panel of L1-specific SNPs to try to characterize the phylogenetic position of L1/SIT109.

The L1/SIT109 sublineage is a clonal complex that is likely to have been introduced in Madagascar long time ago, in relation to peopling. Thus, the transmission of the L1/SIT109 clonal complex occurred in a broad spatial and temporal genetic landscape in Madagascar that remains difficult to decipher, and will tentatively be reconstituted by further WGS studies.

Key worlds: *Mycobacterium tuberculosis* complex, Lineage 1, East-African Indian, MIRU-VNTR, spoligotyping, SNPs, Indian Ocean Trade

INTRODUCTION

Tuberculosis (TB), caused by *Mycobacterium tuberculosis* complex (MTC), remains one of the infectious diseases causing the most deaths worldwide. One third of the world population is infected with *M. tuberculosis*. In 2014, according to WHO, 9.6 million people contracted TB and 1.5 million people died (of which 360.000 HIV-positive) [1].

Actually, molecular genotyping tests have been gradually used in TB transmission studies. This facilitates the identification of the scale of TB transmission (between closely patients or even in a population) [2-4]. These methods also allow to distinguish recent transmission cases, reactivation, latent, or exogenous reinfection cases [5, 6]. Genetic typing techniques such as spoligotyping [7], MIRU-VNTR [8], *IS*6110-RFLP [9] have been developed for TB genotyping. Associated with the geographical distribution data, these methods have been used to classify clinical isolates according to their biogeographical origin [10, 11]. Spoligotyping, Region of Deletions, Whole genome analysis and SNPs, were used to classify *M. tuberculosis* in 7 lineages and many sublineages (L1/EAI, L2/Beijing, L3/CAS, L4/Euro-American, L5/africanum WA1, L6/africanum WA2 and bovis, L7/Ethiopia) [10, 12-14, 15 , 16].

A previous study in Madagascar showed a large genetic diversity of *M. tuberculosis* clinical isolates circulating, a high percentage of the L1/EAI (14%) globally, and a component linked to East Africa (L3/CAS, L4.6.4.2/LAM_ZWE) [17]. The prevalence of L1 was especially high in the coastal provinces of Madagascar and reciprocally L4 clinical isolates seems to predominate in the capital [17]. One predominant and endemic spoligotype, L1/SIT109 and its derivatives, also designated as L1/EAI8_MDG was observed to represent up to 40% on the west coast of Madagascar [17, 18]. The L1/SIT109 is characterized with a spoligotype where spacers 2, 3, 19, 29, 30, 31, 32 and 34 are absent. The L1/SIT109 is found mainly but not exclusively in Madagascar (it is found in Saudi Arabia) and no deeper characterization study was performed on this genotype.

The aim of this study was to analyze the genetic diversity of the Malagasy strain L1/SIT109 using more discriminant genotyping methods and to try to decipher the origin of this MTC genotype in relation to peopling origin, migration and trading routes in the Indian Ocean.

METHODS

Samples:

A total of 156 *M. tuberculosis* clinical isolates previously typed by classical spoligotyping with 43 spacers and harboring the spoligotype L1/SIT109, available in the Mycobacteria unit of the *Institut Pasteur de Madagascar,* were used in the study. Clinical isolates were cultured and identified from Malagasy patients consulting for diagnosis in treatment centers (CDTs) around Madagascar between 1995 and 2010. Clinical isolates were cultured in Löwenstein-Jensen solid medium [19] and DNA was extracted from fresh sub-cultures using the cetyl-trimethyl ammonium bromide (CTAB) method [20].

Study flow:

A first sample of 53 L1/SIT109 clinical isolates were typed with the high throughput spoligotyping method with 68 spacers on a Luminex® 200 system (Luminex Corp. Austin, TX) and by the 24 MIRU-VNTR method for identification of the most discriminant loci [8]. The rest of L1/SIT109 isolates (n=103) were typed using a selection of the 10 most discriminant MIRU-VNTR loci, and finally the diversity of the totality of isolates was analyzed with the selection of loci (Fig.1). The discrimination level of each method and each VNTR locus was computed with the Hunter and Gaston Discriminatory Index (HGDI) method [21] and using online the "Discriminatory Power Calculator" site (http://insilico.ehu.es/mini tools/discriminatory power). A cluster was defined by two or more clinical isolates with identical spoligotypes and the genotypic clusterization rate was defined as the proportion of the sum of clinical isolates with the same profiles. Phylogenetic tree was built online in (*www.miru-vntrplus.org)* using the maximum likehood method.

Spoligotyping:

Amplification:

High-throughput Spoligotyping with 68 spacers on a Luminex 200® was done as described previously on a first sample of 53 L1/SIT109 DNAs [22]. Classical primers designed for spoligotyping described in 1997 were used for amplification of the DR region [7]. The reaction mixture contained 2 µl of a DNA sample (20 to 40 ng), 0.2 mM each deoxynucleoside triphosphate (dNTP), 1 µM of each primer, PCR buffer (10mM Tris-HCl, pH8.3, 50mM KCl), and 1.0U of *Taq* polymerase. The following PCR program was used: 3 min at 95°C, followed by 25 cycles of 30s at 95°C, 30 s at 55°C, and 30 s at 72°C, and a final elongation step at 72°C for 5 min.

Hybridization:

Oligonucleotide-precoupled MicroPlex beads (polystyrene microbeads) were used for hybridization. These reagents (research use only) are available from Beamedex® SAS (www.beamedex.com , Orsay, France).

Hybridization of 2µl of the PCR products with a minimal numbers of 1,800 beads per analyte in 50µl of tetramethylammonium chloride buffer (1X TMAC) was performed after denaturation for 10 min at 95°C and then 20 min at 50°C. After centrifugation at 4,000 rpm and replacement of 35 µl of supernatant by 1X TMAC, streptavidin-phycoerythrin Lumigrade solution (Roche Biochemicals, Meylan, France) prepared in 1X TMAC was added to a final concentration of 2 µg/ml, to reach a final volume of 75 µl. We allowed 5 min of incubation in the system (Luminex® 200 or Magpix) at 50°C before reading the samples.

The Luminex® 200 high-throughput system was used for reading and the xPonent® software (version 3.1.871) was used to analyze the results. Interpretation of results and determination of cut-off were made as in previous study [22].

MIRU-VNTR:

The standard 24 MIRU-VNTR loci method [8] was performed based on agarose gel electrophoresis. The simplex PCR product size was determined as previously reported [23].

L1-Specific Multiplex SNP Analysis

A specific Lineage 1 high-throughput Single Nucleotide Polymorphism (SNP) Typing method was developed by E. Costa Conceicao et al. (results to be published elsewhere). Briefly, this method is a 24-Plex method using 12 DPO primers (dual-priming oligonucleotide) that targets 12 polymorphic SNPs in 12 genes previously shown to be polymorphic [14, 15].The targeted SNPs are: hemL_1104_GA, ftsX_303_GA, moaC1_375_CA, dinP_700_GT, polA_1629_GC, dnaG_51_CG, rv0944_205_CT, rimM_339_CT, rv2707_711_GA, rv3915_1056_GA, glgB_1038_CT, alkA_595_GC.

RESULTS:

Sampling

The Table 1 summarizes the geographic distribution of the collection analyzed in this study. This table shows that L1/SIT109 isolates are present in all provinces of Madagascar. The majority of isolates are found in the capital and in the province of Tulear and the highest proportion of isolates was recovered between 2005 and 2007.

Spoligotyping with 68 spacers:

The genotyping results of a first collection of 53 L1/SIT109 are summarized in Table 2. Four different patterns only were obtained by spoligotyping using 68 spacers. The genotypic clusterisation rate was 94.34%. Three unique profiles and one cluster with 50 isolates were obtained. The HGDI value of a spoligotyping with 68 spacers in this case was 0.111.

24 MIRU-VNTR:

Among the 53 L1/SIT109 clinical isolates, 41 patterns were obtained (Table 2). 36 unique patterns and 5 genotypic clusters with 2 to 7 clinical isolates were identified. The genotypic clusterisation rate was 32.07%. The HGDI value of a 24 MIRU-VNTR is 0.978. The HGDI values of each locus varied from 0 to 0.6103 (Table 2). Nine loci were shown to be invariants within the L1/SIT109 clinical isolates: MIRU02, MIRU20, MIRU24, MIRU26, Mtub04, Mtub29, Mtub30, Mtub34 and Qub4156.

10 MIRU-VNTR:

The ten most discriminant loci observed within the first sample of L1/SIT109 clinical isolates were: ETRA, ETRB, ETRD, MIRU10, ETRE, MIRU40, Mtub21, Mtub39, Qub11b and Qub26. The HGDI values were respectively: 0.2663, 0.5247, 0.4724, 0.2083, 0.2765, 0.3041, 0.4231, 0.6103, 0.1422 and 0.4057. Analysis of all L1/SIT109 clinical isolates (n=156) showed 93 profiles with a clusterization rate of 54.48% (71 single profiles and 22 clusters containing from 2 to 13 isolates). Phylogenetic tree of the 156 L1/SIT109 clinical isolates with this 10 MIRU-VNTR set is shown in Figure 2. The Phylogenetic tree showed a large diversity of clinical isolates in each cluster despite some geographically and temporally clustered cases that were not investigated more deeply.

If we focus on the largest genetic clusters (i.e. with more than two isolates, n=6), the first cluster designated as cluster A (Figure 2) showed 8 isolates recruited from 2005 to 2009. Five of these 8 isolates were from Antananarivo and the remainder cases were found in the 3 provinces of Fianarantsoa, Mahajanga and Tulear. The second genetic cluster (named B) gathers 8 isolates recruited in 2005 and 2006. Two isolates are from Majunga, two from Toamasina, 2 from Tulear and 2 from Fianarantsoa. The third genetic cluster (named C) comprises 9 isolates from 2005 to 2006. Two are from Toamasina, 2 from Tulear, 1 from Antananarivo, 2 from Antsiranana, 1 from Fianarantsoa and 1 from Mahajanga. The fourth genetic cluster (named D) comprised 5 isolates from 2006. Only one of these isolates comes from Antananarivo and 4 come from Tulear. Two isolates very close genotypically to these isolates were also isolated from Tulear at the same period. The fifth genotypic cluster (named E) comprises 13 isolates collected between 2000 and 2010. Eight of these isolates came from Antananarivo, 3 from Fianarantsoa and 2 from Tulear. The sixth genotypic cluster (named F) comprised 7 isolates among which 2 were from Tulear, 2 From Toamasina, 1 from Mahajanga, 1 from Antananarivo and 1 from Fianarantsoa.

The HGDI values of a selection of 10 loci MIRU-VNTR for the totality of clinical isolates was 0.982.

Comparison between spoligotyping 43 spacers, spoligotyping 68 spacers and 24 MIRU-VNTR:

The comparison of HGDI of the different *M. tuberculosis* genotyping methods for SIT109 clinical isolates is reported in the Table 3. Result showed a very less discriminatory power of the spoligotyping with 68 spacers among the SIT109 clinical isolates. The HGDI of the 24 MIRU-VNTR and the selection of 10 MIRU-VNTR is relatively close (respectively 0.978 and 0.970) for the first sample of 61 clinical isolates and demonstrated a high level of diversity.

L1-Specific Multiplex SNPs analysis

During the course of the development of a L1-specific SNPs assay that would be used to distinguish L1 sublineages (E. Costa Conceicão, manuscript in preparation), we genotyped a total of 105 SIT109 clinical isolates to assess their genotype on a panel of 12 L1-specific SNPs. We used as positive controls a set of DNAs belonging to L1.1 (SIT11/EAI3_IND, (India) SIT139/EAI4_VNM (Vietnam), SIT591/EAI6_BGD (Bengladesh), or to L1.2 (SIT48/EAI1_SOM (Somalia, or to SIT19/EAI2_PHL ;Philippines, Manilla type). Even if not all of the 12 genes did allow to get a positive answer on the allelic status of each sample for the time-being, we got robust positive results for 8 genes for most of the samples. The Table 4 summarizes these results. As observed in this table, SIT109 would be a sublineage inside L1.1, since it show to be closer to L1.1/EAI3_IND than to any other sublineage, since we observed only one SNP difference with typical L1.1/EAI3/SIT11, the most frequent L1 type in Tamil Nadu, South India [27], whereas there were at least 2 SNPs difference with the two other positive controls we used inside L1.2 , i.e. EAI2_PHL/SIT19 and EAI1_SOM/SIT48, both typical from L1.2. For the time-being, it is impossible to assign more precisely SIT109 within a more precise phylogenetic position in the L1 phylogeentic tree, and to find the most recent common ancestor to the other L1.1. sublineages; only WGS of this clonal complex will allow to find a likely ancestor and allow to compute the date of divergence with the most recent common ancestor of all L1.1.

Phylogeography and Evolution of L1/SIT109 in Madagascar

In the world-wide TB spoligo-database SITVITWEB, SIT109 is found mainly but not exclusively in Madagascar (56%, n=46), since many identical patterns are found in Saudi Arabia (35%, n=29) and elsewhere (9%, n=7) (See S1_Table). The first description of SIT109 goes back to 1994 in Madagascar. In Saudi Arabia, the origin of patients was mainly saoudian but also found with Indonesian, Afghanistan, and Ethiopian origins. The molecular evolution understanding of SIT109 will deserve whole genome sequencing of various SIT109 samples, however we already performed some hypothesis as suggested by spoligotyping evolution in Figure 3. Figure 3 suggests that L1/SIT236 is the most likely spoligotyping ancestor that allows to get, in two steps, the L1/SIT109 signature. Both L1/SIT126 and L1/SIT2671 are found as likely hypothetical step 1 ancestors in the SITVITWEB database if we hypothesize that loss of spacers 2-3 and loss of spacer 19 are independent events (Figure 3). The phylogeographical specificity of L1/SIT126 (n=77 in SITVITWEB) points to India (41%), Saudi Arabia (11%), Bengladesh (6%), Tanzania (6%), and this genotype is also anecdotally found in Malaysia, Senegal and Uganda (S1_Table). This observation may be refined by considering that SIT126 is the second most prevalent spoligotype cluster in Tamil Nadu, South India (n=80) , just after SIT11 (n=336) in a three year study [27]. Thus, SIT11 and SIT109 might be two independent unique evolutionary events of SIT126. The phylogeographical specificity of the second less likely ancestor, SIT2671 (n=2 in SITVITWEB), is restricted to Saudi Arabia and to one immigrant of unknown origin in the USA.

DISCUSSION

The aim of this study was to evaluate the intra-SIT109 genetic diversity of an historically highly prevalent Lineage 1 subtype in Madagascar, designated as "Malagasy *M. tuberculosis* clonal complex L1/SIT109 or L1/EAI8-MDG, using complementary genotyping methods. Addition of the 25 spacers to the classical 43 spacers in spoligotyping was previously shown to increase the discriminatory level of spoligotyping in L1 and L5-L6 of *M. tuberculosis* complex [24, 25]. We thus decided to use this method. However results were disappointing and showed that the SIT109 remained quite homogeneous with only slight variations. Conversely, as expected, a high level of genetic diversity by the MIRU-VNTR method was observed. The MIRU-VNTR method is known to be more discriminant than spoligotyping when applied to potentially epidemiologically-linked clinical isolates, also providing interesting phylogenetical information [26]. A quite high diversity within the SIT109 clinical isolates was observed in this study. The high level of genetic diversity observed by MIRU-VNTR suggests the historical spread of either a single founding clone (founding effect) or of a limited amount of similar founders clones. It also suggests that SIT109 has been circulating since a long time in Madagascar and had time to evolve.

The geographical distribution of isolates shows that the SIT109 strain is present throughout Madagascar. Assuming that these isolates are derived from a single clone, it may also suggest that the transmission of TB in Madagascar spread on a global scale during Madagascar TB outbreak history. Patients from very remote areas share apparently the same isolates using our methods than patients found in Antananarivo. This also suggests ongoing transmission chains within this clonal complex. TB in Madagascar can therefore be easily transmitted through rapid contacts between tuberculosis patients and healthy people.

Concerning the distribution of clinical isolates inside each genotypic cluster, despite some clinical isolates isolated in the same period and the same region suggesting their recent transmission (5 isolates in Antananarivo inside the cluster A ; 2 isolates in Toamasina, in Mahajanga, in Fianarantsoa et in Tulear inside the cluster B ; 2 isolates from Antsiranana, Tulear an Toamasina in the cluster C ; 4 isolates from Tulear in the cluster D ; 8 isolates from Antananarivo, 3 from Fianarantsoa and 2 from Tulear inside the cluster E ; And 2 isolates from Tulear and 2 isolates from Toamasina inside the cluster F), a large diversity of clinical isolates isolated from different periods and different settings inside each genotypic cluster was observed. This support our hypothesis that the transmission of SIT109 clinical isolates was performed historically with progressive spatial and temporal transmission in some remote regions of Madagascar. This also suggests an homogenous transmission of this strain and probably the same transmission mode of TB in Madagascar.

However, the high prevalence of SIT109 clinical isolates in Madagascar compared to others clinical isolates relatively more virulent such as Beijing clinical isolates suggests that L1/SIT109 isolates are adapted to Malagasy populations, either since these isolates were more transmissible than other clinical isolates due to specific characteristics, or were introduced the most early in TB outbreak history in Madagascar and thus had time to spread and diversify. The co-evolution hypothesis is likely to favor since L1/EAI clinical isolates are known to be less virulent than other *M. tuberculosis* clinical isolates [18, 27, 28].

Previous studies showed characteristic MIRU-VNTR profiles for the EAI family with more than 4 copy number of MIRU23 and more than 1 copy of the MIRU 24 [29]. One strain however had 2 copies of MIRU23 which is in contradiction with this characteristic.

Previous study using 24 MIRU-VNTR showed that all of the 24 loci are variable by considering all of the lineages of *M. tuberculosis* clinical isolates. Except the MIRU02 with a HGDI value of 0.0518, other locus have HGDI values superior to 0.250 [30]. The 9 invariable loci (MIRU03, MIRU20, MIRU24, MIRU26, Mtub04, Mtub29, Mtub30, Mtub34 and Qub4156) are therefore characteristic of the SIT109 clinical isolates.

This study allows to better understand SIT109 strain characteristics circulating in Madagascar and to understand their evolution and their transmission mode. It is also the first subtyping study of one subfamily of *M. tuberculosis* clinical isolates within the EAI lineage.

Even if the Austronesian component is important and could be at the origin of L1/SIT109 in Madagascar, the Bantu component in also ancient and inherent of Madagascar historical peopling, in particular on the West coast of the island [31]. This component was already discussed by Lusitanian sailors during the XVIth century [31]. Thus, the Bantu peopling could equally be responsible of L1/SIT109 introduction in Madagascar. Another variant type of L1/EAI, SIT129 is found to be very prevalent in Mozambique and Malawi, and looks as being more prevalent on the coasts, and could represent a passed common history between Mozambique and Madagascar and could be linked to the historical Indo-Ocean trade (Figure 4) [32]. Arab scholars also showed their important role in population movements in the Mozambique channel, long before European colonization, *i.e*. with an Arab presence in Comoro islands as early as the Xth century [31].

The linguistic approach, pioneered by Otto Ch. Dahl in 1951 shows that Malagasy language and Maanjan have a common origin [33], and that this language points to South-East Kalimantan Barito ethnic group [31]. These people are indigenous ethnic group ; in the 2000 census they made up 2.8% of the Central Kalimantan population [34]. These people are supposed to have migrated to Madagascar island around 945 to 946 AD, sailing through the Indian Ocean on 1,000 leeboard sailboats [31, 35]. .

The recent human genetic study by Hurles *et al*. in 2005 confirms the historical and archeological sources: maternal and paternal heritages are broadly 50-50 between South-East Asian (Borneo) and African components in the Malagasy population. This results does not however facilitate our hypothesis to link L1/SIT109 to Asian or African origins. Based on an hypothetical South Indian L1/SIT126 phylogeographical specificity of L1/SIT109 ancestor, a more recent (XIXth) Indian immigration from South India could also explain the introduction of TB in Madagascar even though it seems that, given the demographical history of Indian communities in Madagascar, and the previous peopling history of the island, such a recent introduction history is less likely than the more ancient south-east Asian or African hypothesis [36].

Timing of migration estimation in relation to SNPs diversity once WGS data will be made available could ultimately shed more light on the African or Asian (Indian or South-East Asian) origin of L1/SIT109. Alternatively, specific geographic microsampling of both *Mycobacterium tuberculosis* and *Homo sapiens* in Mozambique, Borneo and in Madagascar could help to find the link between ancestral Y chromosome - Mt DNA haplotypes and genotypes of MTC that could be at the origin of tuberculosis introduction in Madagascar and more largely on the East African Indian shores.

Some limitation of this study are the lack of more precise clinical epidemiology data and the lack of any estimation time for the introduction of the studied clinical isolates in Madagascar. The second limitation is the poor representativity of samples of the different provinces of Madagascar to analyze more precisely the distribution of the different frequencies of L1/

L1/SIT109 clinical isolates in all regions in Madagascar. And finally a true phylogeny using whole-genome sequencing data is needed to confirm phylogenetic relation between these clinical isolates and try to infer a more precise history of their evolution.

CONCLUSION

Addition of 25 supplementary spacers in spoligotyping was not sufficient to efficiently subtype the L1/SIT109 *M. tuberculosis* clinical isolates. The MIRU-VNTR method was able to discriminate different historically linked and may be epidemiologically linked clusters within L1/SIT109 clinical isolates however VNTR typing with 24 loci is not necessary. Typing with a selection of the 10 most discriminant loci is sufficient to discriminating different clusters within L1/SIT109 clinical isolates with the same discrimination level.

Malagasy L1/SIT109 clinical isolates appear to be quite diverse and can be considered as clinical isolates circulating in Madagascar since many centuries, either in relation to an introduction linked to south-east Asian (Kalimantan) Mannyan population migration during the Xth century or to a Bantu-linked, East-African introduction. Even it seems less likely, we cannot eliminate the introduction of a more recent Indian-linked migration introduction of L1/SIT109 or its ancestor during modern history.

ACKNOWLEDGEMENT

I am grateful to the “Infection Genetics Emerging Pathogens Evolution team” of the Institute of Integrative Cell Biology, UMR9198, CEA-CNRS-University Paris Saclay, France and the Mycobacteria unit of the *Institut Pasteur de Madagascar* for allowing this project to be run, for samples collection, and for financing this project.

BIBLIOGRAPHY

1. WHO: Global Tuberculosis Report 2015. In.: WHO, Geneva, Switzerland; 2015.

2. Bryant JM, Schurch AC, van Deutekom H, Harris SR, de Beer JL, de Jager V, Kremer K, van Hijum SA, Siezen RJ, Borgdorff M *et al*: Inferring patient to patient transmission of *Mycobacterium tuberculosis* from whole genome sequencing data. *BMC Infect Dis* 2013, 13:110.

3. Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L, Brodkin E, Rempel S, Moore R, Zhao Y, Holt R *et al*: Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N Engl J Med* 2011, 364(8):730-739.

4. Walker TM, Monk P, Smith EG, Peto TE: Contact investigations for outbreaks of *Mycobacterium tuberculosis:* advances through whole genome sequencing. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases* 2013, 19(9):796-802.

5. Ford C, Yusim K, Ioerger T, Feng S, Chase M, Greene M, Korber B, Fortune S: *Mycobacterium tuberculosis*-heterogeneity revealed through whole genome sequencing. *Tuberculosis (Edinb)* 2012, 92(3):194-201.

6. Ford CB, Lin PL, Chase MR, Shah RR, Iartchouk O, Galagan J, Mohaideen N, Ioerger TR, Sacchettini JC, Lipsitch M *et al*: Use of whole genome sequencing to estimate the mutation rate of *Mycobacterium tuberculosis* during latent infection. *Nature genetics* 2011, 43(5):482-486.

7. Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper S, Bunschoten A, Molhuizen H, Shaw R, Goyal M *et al*: Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *J Clin Microbiol* 1997, 35(4):907-914.

8. Supply P, Allix C, Lesjean S, Cardoso-Oelemann M, Rusch-Gerdes S, Willery E, Savine E, de Haas P, van Deutekom H, Roring S *et al*: Proposal for Standardization of Optimized Mycobacterial Interspersed Repetitive Unit-Variable-Number Tandem Repeat Typing of *Mycobacterium tuberculosis*. *J Clin Microbiol* 2006, 44(12):4498-4510.

9. Hermans PWM, Solingen DV, Dale JW, Schuitema RJ, Adam RM, Catty D, Embden JDAV: Insertion element IS*986* from *Mycobacterium tuberculosis:* a useful tool for diagnosis and epidemiology of tuberculosis. *J Clin Microbiol* 1990, 28:2051-2085.

10. Brudey K, Driscoll J, Rigouts L, Prodinger WM, Gori A, Al-Hajoj SAM, Allix C, Aristimuno L, Arora J, Baumanis V *et al*: *Mycobacterium tuberculosis* complex genetic diversity : mining the fourth international spoligotyping database (SpolDB4) for classification, Population Genetics, and Epidemiology. *BMC Microbiol* 2006, 6(6):23.

11. Demay C, Liens B, Burguière T, Hill V, Couvin D, Millet J, Mokrousov I, Sola C, Zozio T, Rastogi N: SITVITWEB – A publicly available international multimarker database for studying *Mycobacterium tuberculosis* genetic diversity and molecular epidemiology. *Infect Genet Evol* 2012, 12(4):755-766.

12. Gagneux S, Small PM: Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development. *The Lancet Infectious diseases* 2007, 7(5):328-337.

13. Comas I, Homolka S, Niemann S, Gagneux S: Genotyping of genetically monomorphic bacteria: DNA sequencing in *Mycobacterium tuberculosis* highlights the limitations of current methodologies. *PLoS One* 2009, 4(11):e7815.

14. Coll F, Preston M, Guerra-Assuncao JA, Hill-Cawthorn G, Harris D, Perdigao J, Viveiros M, Portugal I, Drobniewski F, Gagneux S *et al*: PolyTB: A genomic variation map for *Mycobacterium tuberculosis. Tuberculosis (Edinb)* 2014, 94(3):346-54(3):346-354.

15. Coll F, McNerney R, Guerra-Assuncao JA, Glynn JR, Perdigao J, Viveiros M, Portugal I, Pain A, Martin N, Clark TG: A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. *Nature communications* 2014, 5:4812.

16. Stucki D, Brites D, Jeljeli L, Coscolla M, Liu Q, Trauner A, Fenner L, Rutaihwa L, Borrell S, Luo T *et al*: *Mycobacterium tuberculosis* lineage 4 comprises globally distributed and geographically restricted sublineages. *Nature genetics* 2016, 48(12):1535-1543.

17. Ferdinand S, Sola C, Chanteau S, Ramarokoto H, Rasolonavalona T, Rasolofo-Razanamparany V, Rastogi N: A study of spoligotyping-defined *Mycobacterium tuberculosis* clades in relation to the origin of peopling and the demographic history in Madagascar. *Infect Genet Evol* 2005, 5(4):340-348.

18. Rakotosamimanana N, Raharimanga V, Andriamandimby SF, Soares JL, Doherty TM, Ratsitorahina M, Ramarokoto H, Zumla A, Huggett J, Rook G *et al*: Variation in gamma interferon responses to different infecting strains of *Mycobacterium tuberculosis* in acid-fast bacillus smear-positive patients and household contacts in Antananarivo, Madagascar. *Clinical and vaccine immunology : CVI* 2010, 17(7):1094-1103.

19. David H, Levy-Frebault V, Thorel MF: Méthodes de laboratoire pour Mycobactériologie clinique. In., edn. Paris: Institut Pasteur; 1989: 1-87.

20. vanSoolingen D, Hermans PWM, Haas PEWd, Sool DR, Embden JDAv: The occurence and stability of insertion sequences in *Mycobacterium tuberculosis* complex strains: evaluation of an insertion sequence-dependent DNA polymorphism as a tool in the epidemiology of tuberculosis. *J Clin Microbiol* 1991, 29:2578-2586.

21. Hunter PR, Gaston MA: Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. *J ClinMicrobiol* 1988, 26:2465-2466.

22. Zhang J, Abadia E, Refregier G, Tafaj S, Boschiroli ML, Guillard B, Andremont A, Ruimy R, Sola C: *Mycobacterium tuberculosis* complex CRISPR genotyping: improving efficiency, throughput and discriminative power of 'spoligotyping' with new spacers and a microbead-based hybridization assay. *J Med Microbiol* 2010, 59(Pt 3):285-294.

23. Bonura C, Gomgnimbou MK, Refregier G, Aleo A, Fasciana T, Giammanco A, Sola C, Mammina C: Molecular epidemiology of tuberculosis in Sicily, Italy: what has changed after a decade? *BMC Infect Dis* 2014, 14:602.

24. van der Zanden AG, Kremer K, Schouls LM, Caimi K, Cataldi A, Hulleman A, Nagelkerke NJ, van Soolingen D: Improvement of differentiation and interpretability of spoligotyping for *Mycobacterium*

tuberculosis complex isolates by introduction of new spacer oligonucleotides. *J Clin Microbiol* 2002, 40(12):4628-4639.

25. Zhang J, Heng S, Le Moullec S, Refregier G, Gicquel B, Sola C, Guillard B: A first assessment of the genetic diversity of *Mycobacterium tuberculosis* complex in Cambodia. *BMC Infect Dis* 2011, 11(1):42.

26. Bouklata N, Supply P, Jaouhari S, Charof R, Seghrouchni F, Sadki K, El Achhab Y, Nejjari C, Filali-Maltouf A, Lahlou O *et al*: Molecular Typing of *Mycobacterium Tuberculosis* Complex by 24-Locus Based MIRU-VNTR Typing in Conjunction with Spoligotyping to Assess Genetic Diversity of Strains Circulating in Morocco. *PLoS One* 2015, 10(8):e0135695.

27. Narayanan S, Gagneux S, Hari L, Tsolaki AG, Rajasekhar S, Narayanan PR, Small PM, Holmes S, Deriemer K: Genomic interrogation of ancestral *Mycobacterium tuberculosis* from south India. *Infect Genet Evol* 2008, 8(4):474-483.

28. Theus S, Eisenach K, Fomukong N, Silver RF, Cave MD: Beijing family *Mycobacterium tuberculosis* strains differ in their intracellular growth in THP-1 macrophages. *Int J Tuberc Lung Dis* 2007, 11(10):1087-1093.

29. Ferdinand S, Valetudie G, Sola C, Rastogi N: Data mining of *Mycobacterium tuberculosis* complex genotyping results using mycobacterial interspersed repetitive units validates the clonal structure of spoligotyping-defined families. *Res Microbiol* 2004, 155(8):647-654.

30. Devi KR, Bhutia R, Bhowmick S, Mukherjee K, Mahanta J, Narain K: Genetic Diversity of *Mycobacterium tuberculosis* Isolates from Assam, India: Dominance of Beijing Family and Discovery of Two New Clades Related to CAS1_Delhi and EAI Family Based on Spoligotyping and MIRU-VNTR Typing. *PLoS One* 2015, 10(12):e0145860.

31. Allibert C: Migration austronésienne et mise en place de la civilisation malgache. Lectures croisées : linguistique, archéologie, génétique, anthropologieculturelle. *Diogène* 2007, 218:6-17.

32. Sola C, Anselmo LMP, Klotoe B, Panaiotov S, Feliciano C, Costa-Conceiçao E, Namburete EI, Ferro JJ, Bollela VR: On the origin of L1/SIT129 in Beira, Sofala, and on the phylogeography and molecular

evolution of *Mycobacterium tuberculosis* complex in Mozambique and in South-East Africa. *Infection Genetics Evolution* submitted.

33. Dahl OC: Malgache et maanjan : une comparaison linguistique. Oslo: Egede-Institutted; 1951.

34. Ma'anyan people [https://en.wikipedia.org/wiki/Ma%27anyan_people]

35. Hurles ME, Sykes BC, Jobling MA, Forster P: The dual origin of the Malagasy in Island Southeast Asia and East Africa: evidence from maternal and paternal lineages. *American journal of human genetics* 2005, 76(5):894-901.

36. Bardonnet D: Les minorities asiatiques à Madagascar,vol. 10; 1964.

Legend of Figures:

Figure 1: Study flow

Figure 2: Dendrogram of genetic relationships among the 156 L1/SIT109 clinical isolates based on the selection of the 10th most discriminant VNTR loci. The tree was built using neighbor-joining distance algorithm as described previously.

Figure 3: Hypothetical L1/SIT236 and L1/SIT126-based, evolutionary scenario of L1/SIT109 emergence based on two consecutive spacers events; in parallel, other phylogeographical events appeared on other Asian or African countries

Figure 4: Geographical map built using QGIS (www.qgis.org) showing the prevalence of L1-L7 in Madagascar and Mozambique (cf. S2_Table for data source)