Structure génétique de Mycobacterium prototuberculosis et origine de M. tuberculosis

Julien Tap

Master science et technologie
Mention Biologie Moléculaire et Cellulaire
Spécialité Génétique
4 place Jussieu
75252 Paris cedex 05
responsable : M. Sarr

Département infection et épidémiologie
Unité Biodiversité des Bactéries Pathogènes Emmergentes
25-28 rue du Dr Roux
75015 Paris
responsable : M. Brisse

Structure génétique de *Mycobacterium prototuberculosis* et origine de *M. tuberculosis*.

Julien Tap Rapport de stage Mars septembre 2006

Remerciements

Je tiens tous d’abord à remercier Patrick Grimont, responsable de l’unité Biodiversité des Bactéries Pathogènes Emmergentes, de m’avoir accueilli et fait confiance pour ce projet d’étude.

Je présente également mes vifs remerciements à Sylvain Brisse, responsable du laboratoire de génétique des populations bactériennes, pour m’avoir guidé rendant mon stage très enrichissant.

Je remercie Cristina Gutiérrez, chargée de recherche dans le laboratoire de références des mycobactéries, de m’avoir fourni toutes les informations nécessaires pour la réalisation de ce projet.

J’adresse mes remerciements à Virginie Passet, technicienne de l’unité, de m’avoir aiguillé tout au long de mes manipulations *in vitro* et Alexis Deletoile, en thèse dans l’unité, pour ses conseils *in silico*.

Merci enfin à toutes les personnes du laboratoire pour leur accueil dans l’unité.

Résumé

Mycobacterium tuberculosis, l'agent responsable de la tuberculose, est un pathogène mortel de phénotype rugueux infectant un tiers de la population mondiale. *M. tuberculosis* forme avec *M. bovis*, *M. microti*, *M. pinnipedii*, *M. caprae* et *M. africanum* le complexe d'espèce de M. tuberculosis (MTBC). Le MTBC serait issue d'un ancêtre commun vieux de 20000 à 35000 ans. Les études comparatives ont permis de bâtir un scénario évolutif qui unit les membres du MTBC et de révéler que qu'un variant, *M. canettii*, de phénotype lisse serait antérieur à ce complexe clonal.

La découverte et l'isolement de d'autres variants lisses causant la tuberculose a permis, sur l'analyse comparative de séquences de gènes de ménage, de baptiser une espèce progénitrice du MTBC : *Mycobacterium prototuberculosis*. Par ailleurs, Guttierez, et al a mis en évidence des phénomènes de recombinaisons au sein de ces souches lisses.

Sur une population de 56 souches lisses dont 37 souches de *M. cannettii* et 10 membres du MTBC, le séquençage de 16 gènes de ménage a été réalisée. Une analyse des séquences en multi locus a montré l'homogénéité la diversité génétique de *M. prototuberculosis* comparé aux autres espèces bactériennes. Les relations phylogénétiques basées sur ces nouvelles séquences ont confirmés les phénomènes de recombinaison au sein des souches lisses. Dés lors, pour établir des relations de parenté entre les souches en remédiant à la distorsion causée par la recombinaison, un typage multi locus des séquences (MLST) a été réalisé sur l'ensemble de la population. En basant les phylogénies sur les profils allèliques, l'impact relatif de la recombinaison et de la mutation a été évalué chez *M. prototuberculosis*.

Par ces deux approches indépendantes (séquences nucléotidiques et profils allèliques) nous avons montré que *M. prototuberculosis* englobe le MTBC et ces souches lisses au sein d'un groupe homogène et forme l'espèce progénitrice de *M. tuberculosis*.

Mots clés : *Mycobacterium prototuberculosis*, MTBC, *M. tuberculosis*, MLST, phylogénies, recombinaison / mutations.

Sommaire

I. Introduction

A. Les bacilles de la tuberculose

En 1882, le médecin allemand Robert Koch réussit à isoler et à cultiver le bacille responsable de la tuberculose. *Mycobacterium tuberculosis* sera aussi appelé par la suite bacille de Koch. La tuberculose est un problème majeur à l'échelle mondiale. Selon l'Organisation Mondiale de la Santé, près d'un tiers de la population est affecté, avec 1,7 millions de décès en 2004. Une infection due à un bacille tuberculeux a lieu chaque seconde, et de nombreuses souches résistantes aux antibiotiques se propagent. C'est en Asie du Sud-Est et en Afrique subsaharienne, que la tuberculose sévit le plus fortement. Ceci est dû au problème de malnutrition et au VIH qui permet une progression plus rapide de la maladie [23].

Les bacilles tuberculeux sont caractérisés par une croissance lente, une enveloppe cellulaire complexe, une pathogénicité intracellulaire et une homogénéité génétique. Leur temps de génération est de 24 heures. Les mycobactéries font partie de la famille des *Mycobacteriaceae* dans le sous-ordre *Corynebacteriaceae*, ordre des Actinomycetales. Parmi les mycobactéries, il faut distinguer les bacilles causant la tuberculose, comme *M. tuberculosis,* de *M. leprae*, l'agent de la lèpre. De plus, *M. avium*, *M. marinum*, *M. kansasii* et *M. xenopi* sont responsables de mycobactérioses. Les mycobactéries ont une enveloppe cellulaire rugueuse Gram positif avec des peptidoglycanes riches en lipides. Ceux-ci représentent 20 à 45% de l'ensemble de la bactérie, ce qui rend la bactérie peu perméable aux éléments hydrophiles. Parmi ces lipides, l'acide mycolique joue un rôle important dans l'acido-alcoolo-resistance [4].

B. Le complexe d'espèce de Mycobacterium tuberculosis (MTBC)

L'un des buts de la recherche en génétique des populations bactériennes est de comprendre les relations entre d'une part la diversité génétique et les lignées clonales, et d'autre part leurs phénotypes comme la virulence, la transmissibilité, la spécialisation de l'hôte et le succès évolutif [12]. Dans le cas des bacilles tuberculeux, il est nécessaire de s'intéresser en particulier au succès évolutif d'un clone particulier, celui correspondant à *M. tuberculosis* et les autres espèces du complexe d'espèce *M. tuberculosis* (MTBC). Toutes les souches du MTBC sont définies comme étant un unique clone car elles descendent toutes d'une souche ancestrale.

1. Une expansion clonale récente

Le complexe d'espèce *M. tuberculosis* (MTBC) constitue un groupe très compact, et ses membres *M. tuberculosis*, *M. africanum*, *M. bovis*, *M. pinnipedi*, *M. caprae* et *M. microti* peuvent être considérés comme des variants génétiques dérivés de *M. tuberculosis*. Cette homogénéité du MTBC a été établie par hybridation ADN-ADN (>95%), séquençage de l'ARN ribosomique 16S (100% identiques) et séquençage de gènes de codants pour des protéines. L'analyse par séquençage de 26 gènes structurels a montré une faible proportion de substitutions nucléotidiques synonymes chez *M. tuberculosis* (1 nucléotide synonyme variable tous les 10 000 nucléotides synonymes) comparé aux autres bactéries pathogènes, indiquant que la population mondiale de *M. tuberculosis* est issue d'une expansion clonale globale [20]. Le MTBC proviendrait d'un ancêtre commun vieux de 20000 à 35000 ans [20].

Des variations au sein des éléments d'ADN répétés, comme les séquences d'insertion IS6110 et les répétitions directes (DR) ont été trouvées restreintes au MTBC [11], et sont nécessaires pour différencier les souches de *M. tuberculosis* et du MTBC.

2. Historique des scénarios de l'évolution du MTBC

On a longtemps considéré que *M. bovis* était le progéniteur de *M. tuberculosis* , *M. bovis* s'étant adapté à l'homme en donnant *M. tuberculosis* lors de la domestication des bovins. En effet, malgré une préférence pour les bovins, la niche écologique de *M. bovis* est très large. Cette espèce provoque des maladies dans un large panel de mammifères domestiques et sauvages, y compris les humains. Malgré l'homogénéité génétique du MTBC, la niche écologique et la pathogénicité de chaque espèce varient énormément. Le réservoir naturel de *M. tuberculosis* et de *M. africanum* est limité à l'Homme tandis que celui de *M. microti* est limité aux rongeurs [22].

Un autre scénario évolutif de *M. tuberculosis* a été établi par Streevatsan *et al*. Sur la base des séquences de gènes de ménages [20]. Selon ce scénario, le précurseur *M. tuberculosis* est caractérisé par le codon 463 du gène *katG* et le codon 95 du gène *gyrA*. Ces deux sites sont utilisés comme marqueurs génétiques et permettent d'identifier trois groupes phylogénétiques de *M. tuberculosis*. Parmi ces trois groupes, on retrouve la souche de référence *M. tuberculosis* 210 dans le groupe 1, la souche CDC1551 pour le groupe 2 et H37Rv pour le groupe 3. Ce scénario a permis de donner le premier indice que *M. bovis* et *M. tuberculosis* dérivent d'un même ancêtre et non l'un de l'autre.

Les trois génomes de ces souches références ont été séquencés entièrement. L'équipe de Stewart Cole à l'Institut Pasteur a réalisé le séquençage en 1997 de la souche de laboratoire de référence H37Rv. Par

la suite, la souche clinique CDC1551 et *M. tuberculosis* 210 ont aussi été séquencés entièrement à l'institut TIGR par l'équipe de Fleischmann [8]. Le génome de *Mycobacterium tuberculosis* H37Rv a une taille d'environ 4 Mb dont environ 4000 gènes. *M. tuberculosis* diffère des autres génomes bactériens d'une part par sa grande proportion de séquences codantes impliquées dans la production d'enzyme de la lipogenèse et de la lipolyse, et d'autre part par son homogénéité de son taux de G+C%, d'environ 65%. Par ailleurs, chez *M. tuberculosis*, les codons initiateurs de la traduction sont GTG, ATG et ATC [4]. La comparaison par alignement des deux génomes par Hughes *et al*, renforce l'idée d'une forte homogénéité génétique chez *M. tuberculosis* [14].

En 2002, l'analyse de génomique comparative basée sur de la distribution de 20 régions variables résultant d'événements d'insertion/délétion a été effectuée sur 100 souches de *M. tuberculosis*. Cette approche a permis de bâtir un nouveau scénario évolutif pour le MTBC où non seulement *M. tuberculosis* ne proviendrait pas de *M. bovis* mais que le variant *M. canettii*, souche de phénotype lisse isolée en Afrique de l'Est, serait antérieur au précurseur de ce complexe clonal [2].

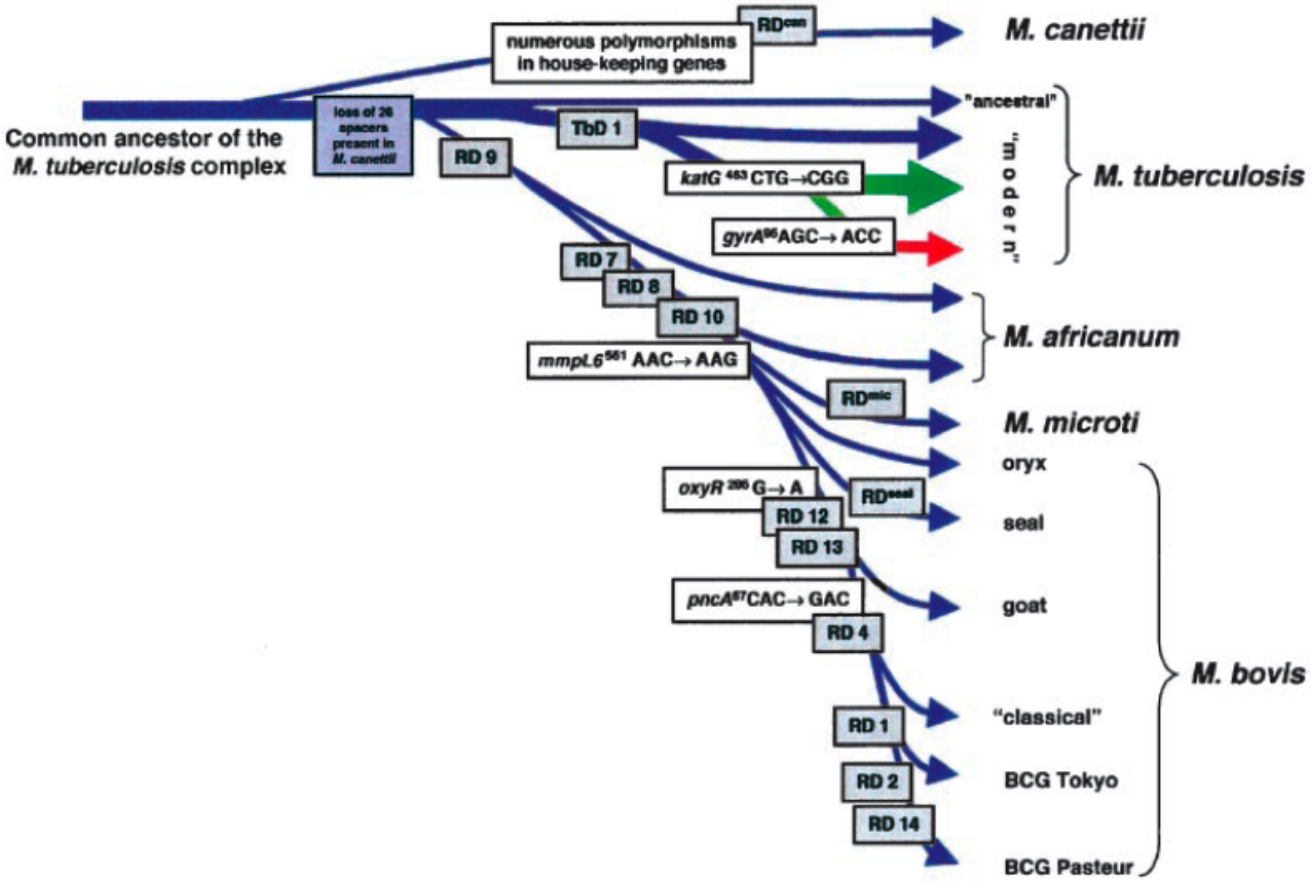


Figure 1: nouveau scénario évolutif proposé par Brosch et al. Ce schéma est basé sur la présence/absence de 'régions de différence' et sur le polymorphisme de 5 gènes sélectionnés. La nouveauté est de proposer que *M. canettii* de phénotype « lisse » est antérieur au précurseur de *M. tuberculosis* [2].

3. Concept d'espèce de *M. prototuberculosis*

Après l'établissement de ce nouveau scénario évolutif, il était intéressant d'orienter les recherches sur l'ancêtre commun de ce complexe d'espèce pour comprendre son origine et son succès évolutif. L'isolement à Djibouti de nouveaux bacilles tuberculeux provoquant la tuberculose mais de phénotype lisse (Fabre *et al* 2004) a permis d'apporter de nouvelles connaissances sur l'histoire évolutive des mycobactéries et en particulier du MTBC [6]. Contrairement à *M. tuberculosis*, ces nouvelles souches forment des colonies lisses comme *M. canettii*, le dernier taxon de bacille de la tuberculose à avoir été décrit [22].

Récemment, des équipes de l'Institut Pasteur (dont une correspondant à l'équipe d'accueil de ce stage) a réalisé une analyse phylogénétique à partir des six gènes de ménages (Guttierez, Brisse *et al*) [13]. Celle-ci montre que le MTBC forme un groupe compact inclu dans un ensemble plus diversifié comprenant les souches lisses (n = 37). L'alignement des séquences concaténées de 3, 387 pb de chaque souches lisses et du MTBC ont permis de mettre en évidence 52 sites polymorphes, dont 46

sites synonymes, un taux nettement supérieur à celui observé dans le MTBC seul. La distance génétique entre le MTBC et les souches lisses est inférieure à celle observée entre des souches lisses [13]. Selon ces nouvelles données, le MTBC est donc un sous-groupe génétique appartenant à un ensemble génétique formé par les souches lisses. MTBC provient donc d'une espèce bactérienne plus diverse et ancienne, baptisée du fait de son antériorité, *M. prototuberculosis* [13].

Ces bacilles tuberculeux lisses ont une origine estimée à 2,6 à 2,8 millions d'années [13] et ils ont été isolés sur des patients atteints de la tuberculose. Par conséquent, il est probable que le dernier ancêtre commun entre ces souches lisses et le MTBC pouvait causer la tuberculose. Ceci suggère que la tuberculose n'est pas récente (plus vieille, par exemple, que la fièvre typhoïde et la malaria). Ces souches lisses ont été isolées en Afrique de l'Est, là où les premiers hommes vécurent. L'opposition entre la diversité génétique de ses souches lisses, mais leur restriction à l'Afrique de l'Est, et l'expansion clonale internationale du MTBC est à mettre en parallèle avec l'hypothèse "OUT OF AFRICA" concernant l'histoire évolutive de l'homme [21].

L'alignement des gènes de ménage chez *M. prototuberculosis* a permis de mettre en évidence une structure de gènes mosaïques, notamment dans *gyrA* et *gyrB*. Ceci montre qu'il y a eu des recombinaisons intra-géniques entre souches lisses. De plus des recombinaisons inter-géniques ont été mises en évidence. Ceci contraste avec l'absence totale de recombinaison observée jusque là dans le MTBC. Cette absence de recombinaison peut être due au fait qu'il ait eu une perte de la capacité des transferts horizontaux, que les événements de transferts horizontaux soient trop rares ou bien que la niche écologique de ces bacilles soit différente, écartant l'opportunité de transferts [13].

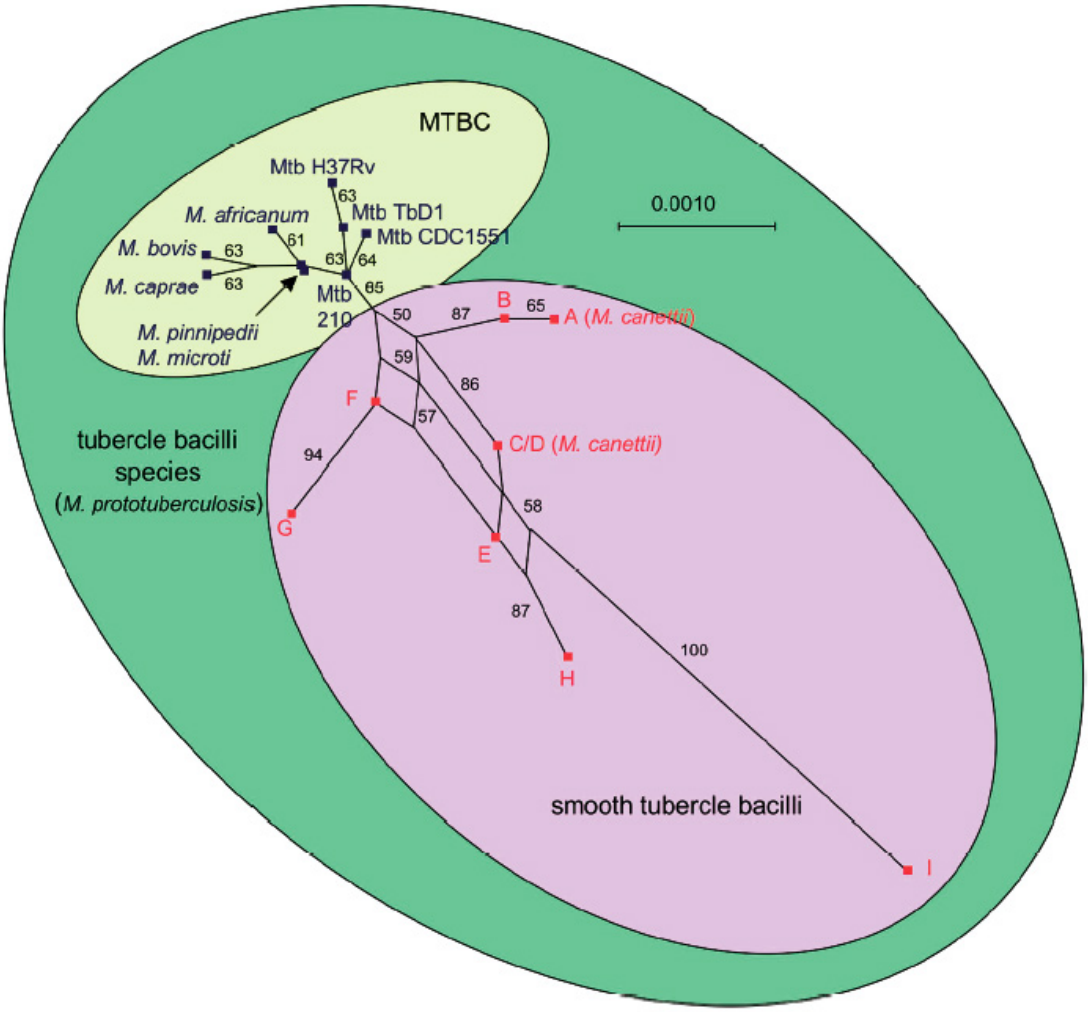


Figure 2 : Split tree réalisée à partir des 17 séquences concaténées des 6 gènes de ménages. Le MTBC de phénotype rugueux forme avec les bacilles tuberculeux lisses une nouvelle espèce baptisée *M. prototuberculosis*. L'échelle représente la distance de Hamming. [13]

C. Objectifs

Le premier objectif de cette étude sera de confirmer que *M. prototuberculosis* forme une espèce regroupant ces souches lisses, y compris *M. canettii*, et le MTBC. Ensuite, le deuxième objectif est d'approfondir notre connaissance de la structure génétique de *M. prototuberculosis*, et en particulier l'importance relative de la recombinaison dans l'évolution des souches. Pour cela, nous séquencerons 16 gènes codant pour des protéines de ménage dans un nombre maximal de souches lisses disponibles (n = 56).

II. Matériels et méthodes

A. Souches bactériennes

Pour cette étude, 66 souches seront analysées 56 souches de bacilles de type lisse et 10 du MTBC. Ces bacilles ont tous été isolés de patient atteint de la tuberculose. (Tableau 1)

Tableau 1 : Souches "lisses" de Mycobacterium étudiées

Code	*Code CNR*	*Espèce*	*Génotype*	*Code secondaire**	*Site d'isolation*	*Date*	*Pays*
M10	*cnr19990160*	*M. canettii*	*CD*		*pulmonary*	*1999*	*France*
M11	*cnr19990161*	*M. canettii*	*CD*		*lymph node*	*1999*	*Djibouti*
M12	*cnr19990263*	*M. prototuberculosis*	*F*	*CIPT140070003*	*pulmonary*	*1997*	*France*
M13	*cnr19990264*	*M. prototuberculosis*	*F*	*CIPT140070012*	*pulmonary*	*1998*	*France*
M14	*cnr19990515*	*M. prototuberculosis*	*B*		*pulmonary*	*1999*	*Djibouti*
M15	*cnr19990516*	*M. canettii*	*CD*		*pulmonary*	*1999*	*Djibouti*
M16	*cnr19990589*	*M. canettii*	*CD*		*pulmonary*	*1999*	*Djibouti*
M17	*cnr19990645*	*M. prototuberculosis*	*L*		*lymph node*	*1997*	*Djibouti*
M2	*cnr19910563*	*ND*	*O*		*nd*	*nd*	*nd*
M21	*cnr19990711*	*M. prototuberculosis*	*B*	*CIPT140070001*	*lymph node*	*1999*	*Djibouti*
M22	*cnr19990768*	*M. prototuberculosis*	*J*	*percy65*	*lymph node*	*1999*	*Djibouti*
M23	*cnr19991574*	*M. canettii*	*CD*		*pulmonary*	*1999*	*France*
M24	*cnr19991669*	*M. prototuberculosis*	*N*		*nd*	*nd*	*nd*
M25	*cnr19991704*	*M. canettii*	*CD*		*pulmonary*	*1999*	*Djibouti*
M26	*cnr19991705*	*M. prototuberculosis*	*H*		*pulmonary*	*1999*	*Djibouti*
M27	*cnr19991708*	*M. canettii*	*CD*		*lymph node*	*1999*	*Djibouti*
M28	*cnr19991709*	*M. prototuberculosis*	*E*	*CIPT140070002*	*lymph node*	*1999*	*Djibouti*
M29	*cnr20000239*	*M. prototuberculosis*	*CD*		*bone*	*1999*	*France*
M3	*cnr19970130*	*M. canettii*	*CD*	*CIPT 140060004*	*pulmonary*	*1997*	*France*
M30	*cnr20000342*	*M. canettii*	*CD*		*nd*	*2000*	*Djibouti*
M31	*cnr20000473*	*M. prototuberculosis*	*I*	*CIPT140070007*	*pulmonary*	*2000*	*Djibouti*
M32	*cnr20000586*	*M. canettii*	*CD*		*pulmonary*	*2000*	*Djibouti*
M33	*cnr20000587*	*M. prototuberculosis*	*G*	*CIPT140070005*	*lymph node*	*2000*	*Djibouti*
M34	*cnr20001049*	*M. canettii*	*CD*		*pulmonary*	*2001*	*Djibouti*
M35	*cnr20001155*	*M. canettii*	*CD*		*pulmonary*	*2000*	*France*
M36	*cnr20001245*	*M. canettii*	*CD*		*pulmonary*	*2000*	*Djibouti*
M37	*cnr20001246*	*M. canettii*	*CD*		*pulmonary*	*2000*	*Djibouti*
M38	*cnr20001247*	*M. canettii*	*CD*		*peritoneal liq.*	*2000*	*Djibouti*
M39	*cnr20001248*	*M. canettii*	*CD*		*pulmonary*	*2000*	*Djibouti*
M4	*cnr19980862*	*M. canettii*	*CD*		*lymph node*	*1998*	*Djibouti*
M40	*cnr20010188*	*M. canettii*	*CD*		*lymph node*	*2001*	*Djibouti*
M41	*cnr20010389*	*M. canettii*	*CD*		*peritoneal liq.*	*2001*	*Djibouti*
M42	*cnr20010390*	*M. canettii*	*CD*		*lymph node*	*2001*	*Djibouti*
M43	*cnr20010391*	*M. prototuberculosis*	*F*		*pulmonary*	*2001*	*Djibouti*
M44	*cnr20010933*	*M. canettii*	*CD*	*CIPT140060017*	*lymph node*	*2001*	*Djibouti*
M45	*cnr20020544*	*M. canettii*	*CD*		*pulmonary*	*2002*	*Djibouti*
M46	*cnr20020986*	*M. canettii*	*CD*		*lymph node*	*2002*	*Djibouti*
M47	*cnr20020987*	*M. canettii*	*CD*		*bone*	*2002*	*Djibouti*
M48	*cnr20020988*	*M. canettii*	*CD*		*pulmonary*	*2002*	*Djibouti*
M49	*cnr20020989*	*M. canettii*	*CD*		*pulmonary*	*2002*	*Djibouti*
M5	*cnr19980863*	*M. prototuberculosis*	*H*	*CIPT 19980863*	*pulmonary*	*1998*	*Djibouti*
M50	*cnr20021261*	*M. canettii*	*CD*		*nd*	*2002*	*Djibouti*
M51	*cnr20030159*	*M. canettii*	*CD*		*pulmonary*	*2002*	*France*
M52	*cnr20030466*	*M. canettii*	*CD*		*lymph node*	*2002*	*Djibouti*
M53	*cnr20030467*	*M. canettii*	*CD*		*blood*	*2003*	*Djibouti*
M54	*cnr20030686*	*M. canettii*	*CD*		*blood*	*2003*	*Djibouti*
M55	*cnr20033147*	*M. canettii*	*CD*		*lymph node*	*2003*	*France*
M56	*cnr20040352*	*M. prototuberculosis*	*M*		*lymph node*	*2003*	*nd*
M57	*cnr20041158*	*M. prototuberculosis*	*K*		*nd*	*nd*	*nd*
M58	*cnr20050462*	*M. prototuberculosis*	*L*		*lymph node*	*1997*	*Djibouti*
M59	*cnr20050642*	*M. canettii*	*CD*		*pulmonary*	*2005*	*Djibouti*
M60	*cnr140010059*	*M. canettii*	*A*	*CIPT140060001*	*pulmonary*	*1969*	*France*
M61	*cnr140010060*	*M. canettii*	*A*		*pulmonary*	*1969*	*France*
M62	*cnr140010061*	*M. canettii*	*A*		*pulmonary*	*1970*	*Papeete*
M8	*cnr19981514*	*M. canettii*	*CD*		*lymph node*	*1998*	*Djibouti*
M9	*cnr19990121*	*M. canettii*	*CD*		*lymph node*	*1993*	*Switzerland*

CIP=Collection Institut Pasteur

Il faut noter que la souche M22 cnr1999076 a été décrite sous le nom de percy65 dans une étude de Fabre *et al* comme une souche divergente. [6] En plus de ce jeu de souches de phénotype lisse, on inclura dans l'étude 10 membres du MTBC dont trois souches de *M. tuberculosis* 210, CDC1551 et H37Rv se situant respectivement dans les groupes 1, 2 et 3 défini par Streevatsan (Voir I.B.2). On inclura dans l'étude une souche atypique de *M. tuberculosis* isolé en Ouganda et une souche de référence de *M. tuberculosis* « ancestrale » (Voir Figure 1) qui contient un locus TBD1 absent des *Mycobacterium tuberculosis* « modernes ».

Tableau 2 : Souches du MTBC incluses dans l'étude

Code	*Code secondaire*	*Espèce*	*Groupe*	*Site d'isolation*	*Date*	*Pays*
Mafricanum		*M. africanum*	*MTBC*			
Mbov_AF2122	*AF2122_97*	*M. bovis*	*MTBC*	*lymph node (cow)*	*1997*	*Royaume-Uni*
Mcaprae		*M. caprae*	*MTBC*			
Mmicroti	*OV254*	*M. microti*	*MTBC*	*voles*	*1930*	*Royaume-Uni*
Mpinnipedii		*M. pinnipedii*	*MTBC*			
Mt_210	*210*	*M. tuberculosis*	*MTBC*	*NA*	*NA*	*Etats-Unis*
Mt_CDC1551	*CDC1551*	*M. tuberculosis*	*MTBC*	*NA*	*NA*	*Etats-Unis*
Mt_H37Rv	*H37Rv*	*M. tuberculosis*	*MTBC*	*human lung*	*1934*	*NA*
MtUganda	*Uganda*	*M. tuberculosis*	*MTBC*			*Ouganda*
TbD1	*TbD1+*	*M. tuberculosis*	*MTBC*			

B. Sélection des locus pour le séquençage

Afin de réaliser la phylogénie de ces souches, il est nécessaire de réaliser le séquençage sur des locus très conservés. Pour analyser les structures de population, il faut étudier les mutations neutres et par conséquent écarter les gènes soumis à sélection comme ceux impliquer dans la structure cellulaire et les gènes de virulences ou de résistance. C'est pourquoi, comme dans l'article de Gutiérrez *et al,* on choisit de séquencer des locus dans des gènes de ménage. [13] En effet, les gènes de ménage sont les gènes indispensables au fonctionnement de l'organisme. Les gènes sont considérés comme structurelle lorsque qui sont impliqués, par exemple, dans des voies métaboliques essentielles. Les gènes de ménages sont sélectionnés en s'appuyant sur l'article de Gil *et al* où est décrit les gènes de ménage pour les bactéries. [10]

Tableau 2 : gènes de ménages sélectionnées

Gènes	*M. tuberculosis* H37Rv	*M. tuberculosis* CDC1551	*M. bovis AF2122_97*	Taille pb	Position en pb sur le génome de H37Rv		Direction
gyrB	Rv0005	MT0005	Mb0005	2145	5123	7267	+
gyrA	Rv0006	MT0006	Mb0006	2517	7302	9818	+
leuS	Rv0041	MT0047	Mb0042	3910	43562	47472	+
hsp65	Rv0440	MT0456	Mb0448	1623	528608	530230	+
rpoB	Rv0667	MT0695	Mb0686	1172	759807	763325	+
adk	Rv0733	MT0757	Mb0754	546	826122	826668	+
pgi	Rv0946c	MT0972	Mb0971c	1662	1055024	1053362	-
katG	Rv1908c	MT1959	Mb1943c	2223	2156111	2153889	-
pncA	Rv2043	MT2103	Mb2069c	561	2288681	2288120	-
glyS	RV2357c	MT2426	Mb2378c	1392	2639673	2638281	-
efp	Rv2534	MT2609	Mb2563c	564	2858727	2858163	-
recA	Rv2737c	MT2806	Mb2756c	2373	3049052	3046679	-
proS	Rv2845	MT2911	Mb2870c	1749	3151202	3149453	-
gltX	Rv2292c	MT3070	Mb3016c	1473	3348805	3347332	-
hpt	Rv3624c	MT3726	Mb3648c	651	4063254	4063904	-
sodA	Rv3846	MT3960	Mb3876	624	4320704	4321327	+

Le typage par séquençage de plusieurs locus de gène de ménage est crucial pour plusieurs raisons. Premièrement, les variations neutres s'accumulent lentement, comparé à la variabilité détectée en électrophorèse en champs pulsés. Le nombre d'allèles dans un seul locus de gènes de ménage est trop faible pour réaliser du typage. C'est pourquoi l'utilisation de plusieurs loci permet de distinguer un nombre plus important de profils alléliques. Plus le nombre de loci de gène de ménages est important plus la probabilité que deux souches différentes possèdent le même profil est réduite voire nulle [19]. De plus pour avoir une bonne représentativité du polymorphisme, il est nécessaire de choisir des gènes repartis sur tous les génomes. En tout, 16 gènes ont été sélectionnés et près de 8kpb (7968pb) seront séquencés par souches. Ceci a pour but de mettre en évidence le polymorphisme dû aux mutations et aux transferts horizontaux au sein de la population.

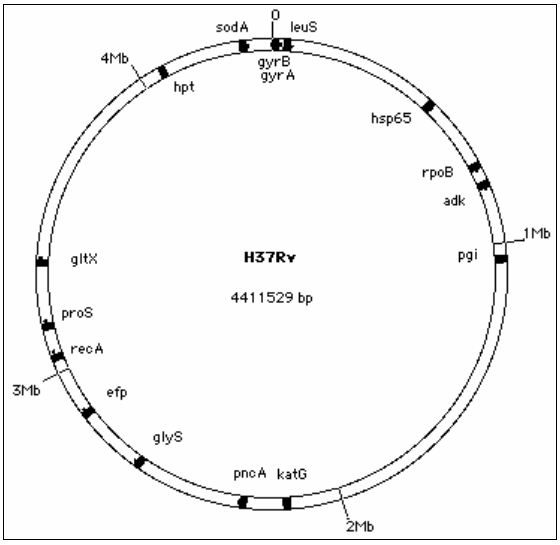


Figure 3: Positions des gènes de ménages sélectionnés sur les génomes de *M. tuberculosis* H37Rv

Les gènes de ménage ont été choisis selon leur répartition sur le génome de H37Rv pour avoir une bonne représentativité du chromosome (Voir Figure 3). Par ailleurs, le choix de certain gène n'est pas

totalement neutre. En effet, beaucoup de protéines avec des motifs répétés « proline- proline-glutamine » ont été mis en évidence dans le génome de *M. tuberculosis*. En tout près de 10% des capacités codantes du génome sont alloués pour ces types de protéines. [4] Il sera donc intéressant de comparer les séquences de ces gènes codant pour la formation de ces acides aminés prépondérants d'un point de vue évolutif.

En ce qui concerne le gène codant pour la catalase peroxydase, *katG*, ce dernier est nécessaire à la survie de la bactérie dans le macrophage en lui permettant de lutter contre le stress oxydatif. Dans l'interaction hôte pathogène, il faut aussi souligner le rôle important du gène *hsp65* qui code pour protéine de résistance et qui est impliqué dans la régulation de la réponse immunitaire en se fixant sur les récepteurs « Toll-like » des macrophages. [3][16]

C. Amorces utilisées

Les nouvelles amorces ont été préalablement dessinées à l'aide du logiciel « primer 3 » en ligne. Les amorces ont été définies par rapport aux séquences codantes pour les gènes de ménages choisis de *M. tuberculosis* H37Rv, disponibles dans la banques de données « TubercuList » (http://genolist.pasteur.fr/Tuberculist/). Pour les gènes *sodA*, *katG*, *gyrA*, *gyrB*, *hsp65* et *rpoB*, on réutilise les mêmes amorces que celles qui ont servies pour l'article de Gutiérrez *et al* [13].

Tableau 3 : couples d'amorces utilisées pour chaque gène

Gène	Nom de l'amorce	Sens	Séquence	Coordonnées (gène)	Taille (b)	Taille de l'amplicon (pb)
leuS2	Rv0041.2F	LEFT	GTCAACCCCTTGTGGACATAC	1105	21	917
	Rv0041.2R	RIGHT	TTGTGCCAGAACCTGGAATAC	2021	21	917
adk	Rv0733F	LEFT	GATCTCCACCGGCGAACTCTT	84	21	462
	Rv0733R	RIGHT	TACTTTCCCAGAGCCCGCAAC	545	21	462
pgi1	Rv0946.1F	LEFT	CGGCGATCTCTACATCGACTA	144	21	892
	Rv0946.1R	RIGHT	CGACAAGTCATTGGAATACGG	1035	21	892
pncA	Rv2043cF	LEFT	GATCATCGTCGACGTGCAGAA	12	21	549
	Rv2043cR	RIGHT	CAGGAGCTGCAAACCAACTCG	560	21	549
glyS	Rv2357cF	LEFT	AGAGAACATCAAGCGCCAGTG	138	21	928
	Rv2357cR	RIGHT	CTTATCCATCCCACCCTTGGT	1065	21	928
efp	Rv2534cF	LEFT	CCACTGCTGACTTCAAGAACG	8	21	522
	Rv2534cR	RIGHT	GCGAATCCACCTTTAGTTTGTC	529	22	522
recA	Rv2737cF	LEFT	CGACAAGATCGGAGTGATGTT	591	21	893
	Rv2737cR	RIGHT	GTTGTTCAGAGGTCGTCGTGT	1483	21	893
proS	Rv2845cF	LEFT	CGCAACATAGAACGGGTCATC	157	21	876
	Rv2845cR	RIGHT	CTTAACCAGGAACGGGTGCTT	1032	21	876
gtlX	Rv2992cF	LEFT	CCCAAGCTGGGTTACGACAAT	382	21	888
	Rv2992cR	RIGHT	CCAGTCCGTCACACTTGTCAG	1269	21	888
gyrA	GyrA-F	LEFT	GTTCGTGTGTTGCGTCAAGT	992	20	1013
	GyrA-R	RIGHT	CAGCTGGGTGTGCTTGTAAA	2005	20	1013
katG	KatG-F	LEFT	CTACCAGCACCGTCATCTCA	953	20	913
	KatG-R	RIGHT	AGGTCGTATGGACGAACACC	1866	20	913
gyrB	MTUBf	LEFT	TCGGACGCGTATGCGATATC	448	20	1020
	MTUBr	RIGHT	ACATACAGTTCGGACTTGCG	1468	20	1020
rpoB	Rpo3'	LEFT	GGATGTTGATCAGGGTCTGC	884	20	340
	Rpo5'	RIGHT	TCAAGGAGAAGCGCTACGA	1224	20	340
hsp65	Tb11	LEFT	ACCAACGATGGTGTGTCCAT	145	20	421
	Tb12	RIGHT	CTTGTCGAACCGCATACCCT	566	20	421
sodA	sodAZ205	LEFT	AGCTTCACCACAGCAAGCACCA	77	22	465
	sodAZ212	RIGHT	GCCCAGTTCACGACGTTCCAAA	542	22	465

D. Amplification des locus par PCR

Après quelques tests préliminaires, chaque locus est amplifié par PCR à partir de chaque ADN génomique des 56 souches lisses et 5 souches du MTBC (*M .caprae*, *M. pinnipedi*, *M. tuberculosis* TbD1+, *M. tuberculosis* Ouganda, *M. africanum*) du Tableau 1. Le mélange réactionnel de PCR de 50 µL est composé de 2 µL d'ADN 10ng/µL, 1 µM de la paire d'amorce, 0,85 U de Taq polymérase Invitrogen, du MgCl2 à 1,5 mM, un mélange des quatre dNTP à 0,2 mM et du tampon à 1X. Les conditions de PCR sont présentées dans le tableau suivant :

Tableau 4 : conditions de la PCR

Température	Temps	Cycles
94°C	4min	1X
94°C	1min	
57°C	1min	30X
72°C	**	
72°C	7 min	1X

*1min d'élongation pour les produits inférieurs à 600 pb, 1min30 pour les autres

Chaque réaction PCR est vérifiée avec une migration sur gel (agarose 1%, tris borate EDTA 1X, bromure d'éthidium) par électrophorèse. Chaque produit PCR est révélé sous U.V. après une migration de 30 min à 80V. Tous les amplicons ont bien la taille attendue. Chaque produit PCR est ensuite purifié par filtration avec 60 µL d'eau stérile. Les produits PCR sont resuspendus par la suite dans 50 µL.

E. Séquençage des produits PCR et analyse des chromatogrammes

Les amorces utilisées pour le séquençage sont les mêmes que pour l'amplification par PCR. Le séquençage se déroule en trois étapes. Tous d'abord la réaction de séquence qui permet de marquer les nucléotides par les fluorochromes en utilisant le trousseau Big Dye Terminator Cycle Sequencing Ready Reaction version 3.1 (perkin-Elmer), puis une purification est réalisée et enfin une chromatographie est effectuée sur un séquenceur automatique à capillaire de type ABI-3700.

Tableau 5 : Mélange réactionnel pour une réaction de séquence

Réactifs	Volume
Produit PCR purifié	2 µL
Tampon 5X	1,5 µL
Amorce (1pmol/L)	3 µL
BigDye	1 µL
H2O	2,5 µL

Tableau 6 : Cycle de la réaction de séquence

Température	Temps	Cycles
96°C	10 s	25X
50°C	5 s	
60°C	4min	

La purification de séquences s'effectue par précipitation avec pour chaque réaction 1 µL d'acétate de sodium 3M, 1 µL d'EDTA 125 mM et 50 µL d'éthanol 95%. Après lavage à l'éthanol 70% et séchage, les plaques peuvent être conservé à -20°C.

L'analyse de chaque chromatogramme est réalisée avec le logiciel BioNumerics. Une matrice de lecture de la séquence est définie pour chaque gène afin que la taille de la séquence étudiée soit la même pour toutes les souches. Chaque base doit être soutenu au minimum par deux chromatogrammes (sens et anti-sens). Pour chaque matrice construite, le cadre de lecture est vérifié en traduisant et en réalisant un BLAST de la séquence protéique obtenue sur les banques publiques.

Tableau 7 : Positions des matrices d'édition en fonction des gènes

Gène	Description	Taille du gène (pb)	Position de la matrice sur H37Rv (pb)	Position sur le gène (pb)	Taille de la séquence éditée (pb)
adk	Adenylate cyclase	546	826317-826592	196-471	276
efp	Elongation factor P	564	2859200-2858889	91-402	312
gltX	Glutamyl tRNA transferase	1473	3349770-3349182	508-1096	588
glyS	Glycyl tRNA synthase	1392	2640761-2640222	304-855	552
GyrA	Gyrase sous unité A	2517	7521-8261	220-960	741
GyrB	Gyrase sous unité B	2145	5624-6559	502-1437	936
hpt	Hypoxanthine guanine phosphoribosyltransferase	615	4063643-4063392	262-513	252
Hsp65/groEL2	Heat shock protein 65	1623	52880-529171	193-564	372
katG	Catalase peroxidase	2223	2154812-2154261	1300-1851	552
leuS	Leucyl-ARNt transferase	2910	44760-45448	1199-1887	687
pgi	Glucose 6 phosphate isomérase	1662	1056436-1055896	250-814	564
pncA	Pyrazinamidase nicotamidase	561	2289142-2288767	100-475	375
proS	Prolyl tRNA synthase	1749	3152693-3152042	258-909	651
recA	Recombinase A	2373	3050641-3050154	784-1271	489
rpoB	sous-unité beta ARN polymerase	3519	760704-761012	898-1206	309
sodA	Superoxide dismutase	624	4320908-4321219	205-516	312

F. Traitements informatiques et statistiques des séquences

1. Quantification de la diversité et analyses phylogénétiques

Pour chaque locus sur l'ensemble des isolats, les paramètres de diversités nucléotidiques (Ks, Ka, Pi) seront calculés avec DNAsp :

-Pi : l'indice de diversité de Nei

-Ks : Le nombre de substitutions synonymes par sites synonymes

-Ka : le nombre de substitutions non synonymes par sites non-synonymes

Sur *bionumerics*, un alignement primaire de types UPGMA sera réalisé pour chaque locus afin de mettre en évidence les différents génotypes au sein des souches de phénotypes lisses. Le nombre de mutation et le pourcentage d'homologie par rapport à la souche de référence H37Rv sera calculé pour chaque groupe et chaque variant du complexe d'espèces de *M. tuberculosis*.

Un alignement multiple est réalisé avec *clustalW* pour toutes les séquences de chaque groupe. Ensuite, on calcule une matrice de distance en Neighbourg Joining et les coordonnées d'un arbre phylogénétique pour chaque locus. Ceci permettra de comparer la congruence de chaque arbre pour chaque locus et de connaître la distance maximale. Un arbre consensus en Neighbourg Joining (NJ) sera généré par *bionumerics* à partir de l'ensemble des locus. Ainsi qu'un arbre en NJ par les séquences concaténés pour chaque groupe.

2. Recombinaison

Afin, de mettre en évidence les phénomènes de recombinaison et les complexes clonaux, la représentation en sous forme d'un arbre en réseaux (split tree) sera réalisées a partir des séquences concaténés pour chaque groupe avec le logiciel *split tree 3.2*.

3. Structuration de la population

Pour finir, un sequençotypage multilocus (MLST, multi locus sequence typing) sera réalisé. Le principe de cette méthode d'analyse consiste à assigner un numéro d'allèle à chaque variant allèlique. Deux souches ayant la même séquence pour un des gènes auront le même numéro pour le locus donné. A chaque nouvel allèle, un nouveau numéro d'allèle est attribué au locus. Le sequençotype (profil allèlique) de chaque isolat sera défini par 16 chiffres. L'ensemble de ces sequençotypes vont constituer une base de données pour l'analyse MLST.

Cette base de données MLST constituée permettra par la suite de réaliser un arbre minimal de chevauchement (minimal spanning tree) grâce au logiciel BioNumerics. Cet arbre permettra de visualiser la structuration de la population.

III. Résultats et discussion

A. Type de variation des gènes analysés

Afin de poursuivre l'analyse de la diversité génétique des souches lisses de bacille de la tuberculose, nous avons analysé les 56 souches lisses disponibles dans nos laboratoires avec les 6 gènes analysés précédemment sur un sous-ensemble de 37 souches lisses (Gutiérrez et al.) et 10 gènes nouvellement sélectionnés. Au total, les 16 gènes ont pu être amplifiés par PCR sur toutes les souches, sauf quelques exceptions pour lesquelles les PCR ont été négatives malgré nos essais répétés avec différentes conditions expérimentales: *hsp65* et *leuS* pour la souche M24, *proS* pour les souches M2, *M. tuberculosis* souche Ouganda et *M. tuberculosis* souche TbD1+, et *recA* pour les souches M22 et M57.

Parmi les souches analysées, la souche M2 présente des caractéristiques atypiques. Celle-ci est très divergente pour les gènes *gyrA* (86,3% de similitude avec les souches les plus proches, du groupe CD), *hsp65* (91,5%), *rpoB* (90,1%) et *sodA* (82%). Pour les autres gènes, cette souche est identique aux souches du groupe CD (voir plus loin ; le positionnement de cette souche dans les analyses phylogénétiques est donné en Annexe). On peut établir deux hypothèses pour cette souche : soit l'échantillon a été contaminé et la séquence des gènes atypiques correspond au contaminant ; soit cette souche a reçu ces gènes atypiques d'un donneur éloigné. Les mutations étant réparties uniformément le long de la séquence divergente, nous n'avons pas vu d'évidence de séquence mosaïque, mais les extrémités du fragment recombiné pourraient être extérieures à la zone séquencée. Comme nous ne pouvons exclure un cas de contamination ou mélange de souches, il nous paraît préférable d'écarter cette souche atypique des résultats présentés. Dans les deux hypothèses, les séquences atypiques ne correspondent à aucune espèce de mycobactérie connue (divergence minimale de 8%).

La longueur des fragments analysés a varié entre 276 paires de bases (pb) pour le gène *adk* et 936 pb pour le gène *gyrB* (voir Tableau 7). Les alignements des séquences n'ont révélé aucune insertion-déletion. Comme observé de manière classique pour les gènes codant pour des protéines de ménage, le nombre de substitutions synonymes est supérieur au nombre de substitutions non synonymes (excepté dans le gène *proS)*. Par ailleurs, le nombre de substitutions synonymes par site synonyme (Ks) est supérieur au nombre de substitutions non synonymes par site non synonyme (Ka). En excluant *proS*, le rapport Ks/Ka varie entre 5,29 pour *pncA* et 111 pour *gltX*. Les changements d'acides aminés sont donc contre-sélectionnés sur ces protéines.

Des études récentes ont montré que le rapport Ks/Ka (ω) varie en fonction du temps évolutif pour les génomes (Rocha et al. 2006). Lorsque que l'on compare le ω des souches du MTBC seules et des souches lisses seules (Tableau 8), on remarque que le ω moyen est différent (1,67% et 19,12%, respectivement). Cette différence est concordante avec l'hypothèse de la récente expansion clonale du MTBC, et un plus long temps évolutif séparant les souches lisses.

Tableau 7 : Polymorphismes des 16 gènes analysés

Gènes	Longueur	N°Site Synonymes	N°Site Non Syn.	N°Sub/Site Synonymes	N°Sub/Site Non-Synonymes	Ks	Ka	ω
adk	276	68,85	207,15	7	0	0,00451	0	>22,5*
efp	312	80,01	231,99	4	1	0,00701	0,00013	52,631
gltX	588	148,56	439,44	10	2	0,01577	0,00014	111,111
glyS	552	123	429	6	3	0,00609	0,00021	29,411
gyrA	741	190,33	550,67	16	1	0,02387	0,00006	333,33
gyrB	936	224,35	711,65	17	3	0,01305	0,00029	45,45
hpt	252	70	182	9	0	0,01792	0	>89,6*
hsp65	372	94,92	277,08	6	1	0,00673	0,00011	62,5
katG	552	140	412	5	2	0,00782	0,00029	27,027
leuS	687	173,03	513,97	15	5	0,00676	0,00052	12,987
pgi	564	146,89	417,11	18	5	0,00853	0,00131	6,493
pncA	375	95,35	279,65	5	6	0,00718	0,00136	5,291
proS	651	171,82	0,00146	6	7	0,00129	0,00146	0,883
recA	489	130,61	358,39	8	3	0,01305	0,00141	9,259
rpoB	309	80,69	228,31	7	4	0,00556	0,0008	6,944
sodA	312	75,02	236,98	7	1	0,00676	0,00026	26,31

* estimation de Ks/Ka avec au moins une mutation non synonymes.

En reprenant le cas particulier du gène *proS* pour lequel ω, était seulement de 0.88%, on remarque que les variations non-synonymes sont essentiellement observées dans le MTBC. Ceci peut résulter du fait de l'expansion récente de ce complexe, la sélection naturelle n'a pas encore eu le temps d'éliminer ces mutations légèrement délétères. Alternativement, il se pourrait que ces mutations non-synonymes présentent un avantage sélectif.

Tableau 8 : Moyenne en % des distances aux sites de substitutions synonymes (Ks) et non synonymes (Ka)

Gène (Taille)	Ks MTBC	Ka MTBC	Ks Souches Lisses	Ka Souches Lisses
adk (276pb)	0,522	0	0,428	0
efp (312pb)	0	0	0,618	0,015
gltX (588pb)	0,135	0	0,986	0,016
glyS (552pb)	0	0,047	0,532	0,017
gyrA (741pb)	0	0,036	1,957	0
gyrB (936pb)	0,426	0,106	1,351	0,031
hpt (252pb)	0	0	2,080	0
hsp65 (372pb)	0,210	0,072	0,416	0
katG (552pb)	0	0,113	0,894	0,009
leuS (687pb)	0,116	0,078	0,771	0,047
pgi (564pb)	0	0	0,799	0,153
pncA (375pb)	0	0,072	0,517	0,147
proS (651pb)	0	0,142	0,148	0,118
recA (489pb)	0	0	0,887	0,078
rpoB (309pb)	0	0,175	0,646	0,062
sodA (312pb)	0	0	0,794	0,030

B. Diversité des génotypes

Le nombre d'allèles rencontrés par gène dans la population des 56 souches analysées varie de 4 (*adk*) à 12 (*gyrB*). La combinaison des 16 gènes pour chaque souche constitue le profil allélique, ou génotype, de chaque souche. Au total, 14 génotypes ont été distingués, nommés de A à O ; 13 en excluant la souche M2 (qui représente le génotype O). Gutteriez et al. (2005) ont défini 9 groupes sur la base de gènes répétés ou d'insertions/délétions génomiques (séquences d'insertions, locus DR, régions de différences)). Les groupes A, C et D correspondent à l'espèce *M. canettii*, le groupe B est très proche de *M. canettii* mais marqué par la présence de la région de différence 12 (RD12) et l'absence de séquence d'insertion IS1080. Cinq autres groupes (E à I) sont plus divergents, sur la base de ces marqueurs, de *M. canettii* et du MTBC.

Un ou deux représentants de chaque groupe avaient été séquencé pour les 9 groupes ainsi identifiés [13]. La comparaison des séquences ne permettait pas de distinguer le groupe C du groupe D. Dans notre étude, nous avons séquencé toutes les souches disponibles pour chaque groupe et confirmons l'absence totale de variation nucléotidique au sein de ces groupes, qui peuvent donc être considérés comme des génotypes (ou séquençotypes, c'est-à-dire génotype défini sur la base de séquences). Les groupes C et D ne montrent pas de variation nucléotidique et peuvent être considérés comme un seul et même génotype, CD. Dans notre étude, 6 nouveaux génotypes (J à O) ont été identifiés (Tableau 1). Sur les 56 souches lisses, 36 (64%) sont de génotype CD, le groupe majoritaire de la population. Les autres génotypes représentés par plus d'une souche sont A (n = 3 souches), F (n = 3), B (n = 2), L (n = 2) et H (n = 2). Les six autres génotypes (E, G, I, J, K, M, N, O) n'ont qu'un seul représentant parmi les souches lisses.

C. Quantification de la diversité génétique et comparaison avec d'autres espèces

Le Tableau 9 présente le polymorphisme au sein de la population en excluant la souche M2. Le pourcentage de sites variables varie entre 1,60% pour le gene *efp* et 19,6% pour *sodA*. Le gène *efp* a également été trouvé le plus conservé parmi 6 gènes séquencés chez *Acinetobacter* (Ecker et al 2006).

Tableau 9 : Variation et diversité des 16 gènes analysés pour l'ensemble de la population (55 souches lisses et 10 souches du MTBC)

Gènes	Longueur éditée (pb)	Sites variables	% Sites polymorphes	Singletons	Sites Informatifs	Allèles	π (%)±ecart type
adk	276	7	2,54%	4	3	4	0,109 (± 0,062)
efp	312	5	1,60%	4	1	5	0,189 (± 0,027)
gltX	588	12	2,04%	4	8	7	0,4 (± 0,056)
glyS	552	9	1,63%	6	3	7	0,152 (± 0,036)
gyrA	741	17	2,35%	4	13	8	0,617 (± 0,161)
gyrB	936	20	2,14%	6	14	12	0,335 (± 0,047)
hpt	252	9	3,57%	1	9	4	0,499 (± 0,061)
hsp65	372	7	1,88%	5	2	6	0,18 (± 0,117)
katG	552	7	1,27%	2	5	5	0,22 (± 0,076)
leuS	687	20	2,91%	11	9	10	0,209 (± 0,038)
pgi	564	23	4,08%	11	12	6	0,318 (± 0,064)
pncA	375	11	2,93%	2	9	7	0,284 (± 0,076)
proS	651	13	2,00%	10	3	6	0,142 (± 0,049)
recA	489	11	2,25%	1	10	6	0,452 (± 0,051)
rpoB	309	11	3,56%	9	2	6	0,204 (± 0,035)
sodA	312	8	2,56%	1	7	4	0,182 (± 0,066)

Le taux de sites polymorphes, le nombre d'allèles et le nombre de sites informatifs d'un point de vue phylogénétique (polymorphisme présent ou absent dans au moins deux souches) sont variables en fonction du gène séquencé.

Pour estimer le polymorphisme de l'ADN, on utilise l'indice de Nei (π) qui est le taux de différences nucléotidiques par site entre deux séquences. Cet indice est défini par l'équation suivante :

$$p = \sum_{ij}^{q} x_i x_j d_{ij}$$

où q est le nombre total d'allèles, x_i est la fréquence de l'allèle i dans la population et d_{ij} est le nombre de différences nucléotidiques ou substitutions par site entre les allèles i et j. En considérant les souches lisses et les membres du MTBC ensemble, cet indice de diversité nucléotidique π varie entre 0,109 % pour le gène *adk* et 1,021 % pour le gène *gyrA*.

Le paramètre π donne un ordre de grandeur de la diversité de la population. Cependant le paramètre π pose problème car la surreprésentation du génotype CD (n = 36) induit une diminution de la diversité nucléotidique. Afin de contourner ce biais, nous avons estimé la diversité π en ne prenant qu'un seul individu par génotype (Tableau 10). Comme attendu, ce π sur les génotypes est plus grand que sur les individus, et l'effet est plus fort sur les gènes les moins variables.

Dans le cadre du débat sur le concept d'espèce chez les bactéries, nous avons cherché à comparer la diversité nucléotidique de *M. prototuberculosis* avec celle des autres espèces bactériennes pour lesquelles des données de même type (MLST) sont disponibles. Les bases de données MLST contiennent pour plusieurs gènes (en général 7) tous les allèles trouvés dans une espèce donnée.

Cependant la fréquence des allèles et génotypes dans les populations naturelles n'est pas directement disponible dans ces bases. En conséquence, nous avons comparé la diversité nucléotidique π entre les allèles distincts, pour chaque gène. Par exemple, la comparaison de séquences pour gène *pgi* permet de distinguer 6 allèles chez *M. prototuberculosis*, on effectuera donc le calcul de π pour ce gène dans cette espèce sur la comparaison de 6 séquences (1 séquence par allèle différent). Le Tableau 10 montre que, comme attendu, le π calculé sur les allèles (qui représente la probabilité de tirer deux nucléotides différents parmi les allèles distincts) est nettement supérieur au π de la population réelle ou des génotypes.

Tableau 10 : diversités nucléotidiques (p **) au sein des individus, des génotypes et des allèles**

		π (%)	
Gènes	Individus	Génotypes	Allèles
adk	0,109	0,294	1,389
efp	0,189	0,217	0,705
gltX	0,4	0,537	0,81
glyS	0,152	0,261	0,552
gyrA	0,617	0,635	1,024
gyrB	0,335	0,517	0,575
hpt	0,499	0,694	1,984
hsp65	0,180	0,363	0,691
katG	0,22	0,361	0,616
leuS	0,209	0,394	0,825
pgi	0,318	0,574	1,726
pncA	0,284	0,567	1,049
proS	0,142	0,249	0,768
recA	0,452	0,713	0,886
rpoB	0,204	0,396	1,251
sodA	0,182	0,466	1,335

Le même calcul du π , en fonction des séquences des allèles obtenus sur les bases MLST, a été réalisé pour les espèces *Listeria monocytogenes*, *Bacillus cereus*, *Escherichia coli* et *Salmonella enterica*. Ces π sont reportés sur la Figure 4 pour chaque gène analysé de ces espèces. On note que comme chez *M. prototuberculosis*, la diversité nucléotidique varie en fonction des gènes au sein d'une même espèce. Bien que les gènes séquencés ne soient pas les mêmes d'une espèce à l'autre, la comparaison globale de ces π « alléliques » montre que les souches analysées de *M. protuberculosis* forment un groupe plus homogène que les espèces comparatives. Le π minimum dans les trois espèces comparatives est observé pour le gène *purA* de *E. coli* (π = 1,50%) et le π maximum pour le gène *ilv* chez *B. cereus* (π = 7,92%). Par comparaison, chez *M. prototuberculosis* les π vont de 0,492 % à 1,984 %. De ce point de vue purement quantitatif, il n'est donc pas aberrant de considérer l'ensemble des souches lisses et le MTBC comme appartenant à une seule et même espèce biologique.

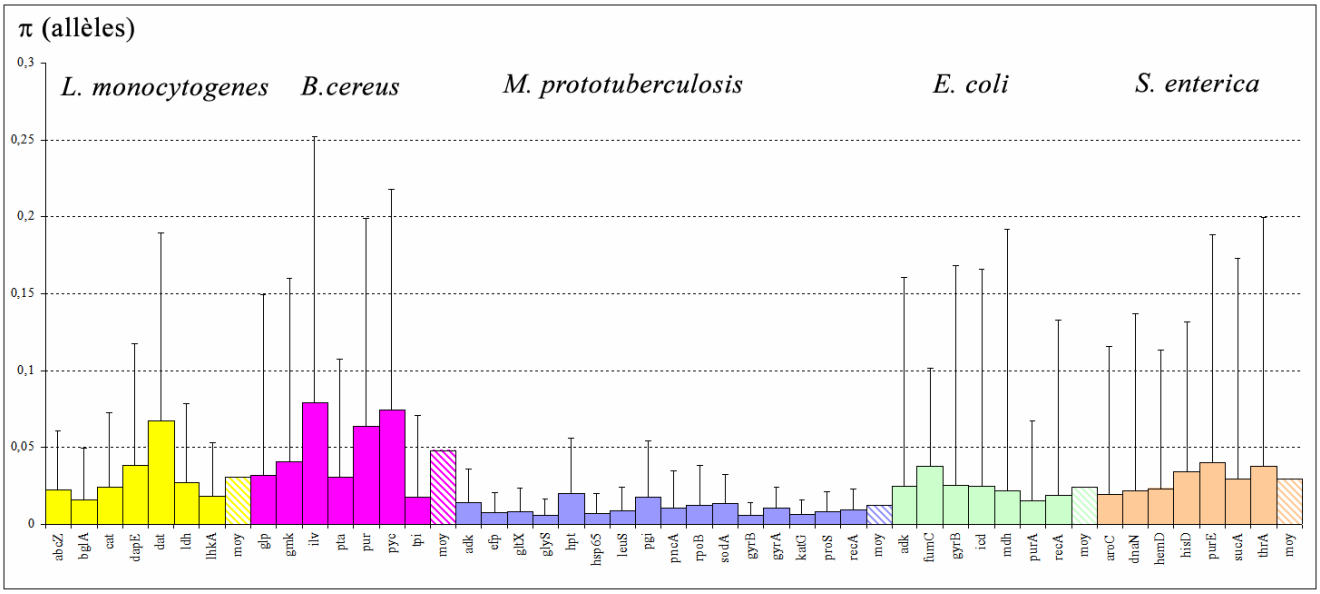


Figure 4 : Diversité nucléotidique entre allèles. Les histogrammes représentent le p **allélique pour chaque gène et la moyenne pour l'espèce (moy, hachuré). Les barres représentent la distance maximale entre deux allèles (et non l'écart-type).**

D. Relations phylogénétiques

La Figure 5 représente les relations phylogénétiques entre génotypes basées sur les séquences des gènes *glyS* et *rpoB*. Les distances maximales entre séquences sont de 1,1% pour le gène *glyS* et de 2,3% pour le gène *rpoB.* Dans les deux cas, le génotype J (représenté par la souche M22, alias percy65) est le plus divergent. On peut remarquer que le génotype F est groupé avec K avec le gène *glyS*, mais est groupé avec le génotype CD avec le gène *rpoB*. Les arbres sont donc incongruents, ce qui peut être expliqué par de la recombinaison génétique au sein des souches lisses. On peut trouver les arbres réalisés à partir des séquences des autres gènes en annexes. D'autres exemples très nets d'incongruence sont visibles. Malgré la présence évidente de recombinaison inter-génique attestée par les incongruences, nous n'avons pas détecté de séquences mosaïques, excepté pour le cas des gènes *gyrB et gyrA* comme déjà décrit précédemment (Guttierez et al. 2005).

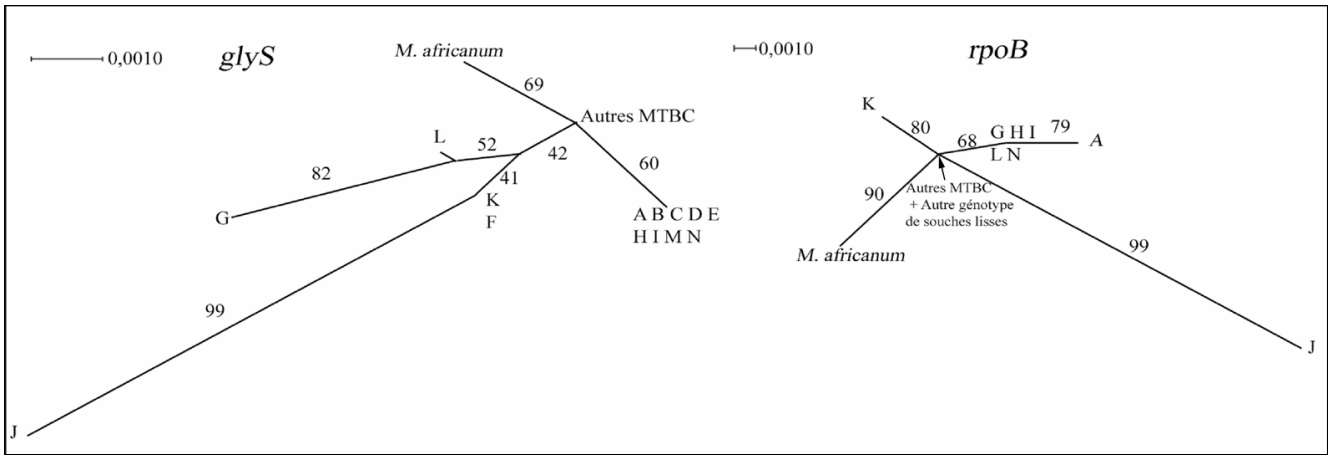


Figure 5 : Phylogénies non enracinée des gènes *glyS* et rpoB basées sur 14 génotypes de souches lisses et les membres du MTBC. L'échelle représente la distance p (non corrigée). Les valeurs de bootstrap sont indiquées sur les branches.

Afin de déterminer les relations phylogénétiques les plus probables qui unissent les différents génotypes de souches lisses entre eux et avec ceux du MTBC, il est nécessaire de s'affranchir des inconguences liés à la recombinaison. Pour cela, on utilise la séquence concaténée de 12 gènes disponibles pour toutes les souches (*adk*, *efp*, *gltX*, *glyS*, *gyrA*, *gyrB*, *hpt*, *katG*, *pgi*, *pncA*, *rpoB*, *sodA*) pour construire la phylogénie. La Figure 6 représente la phylogénie des 14 génotypes de souches lisses avec les différents membres du MTBC. L'ensemble forme un groupe compact où la distance maximale est de 1,5% entre le génotype K et J. On peut distinguer deux groupes de souches lisses soutenus par des bootstraps supérieurs à 67 % : un groupe formé des génotypes G, L, M, N et un autre formé des génotypes A, B, C, D, E. Les autres génotypes (F, I, J, H, K) sont plus divergents et ne paraissent pas fortement apparentés. Les souches du MTBC forment un ensemble compact, étant regroupées avec des distances inférieures à 0,1 %. Les branches qui relient tous les groupes sont très peu robustes (faibles valeurs de bootstraps).

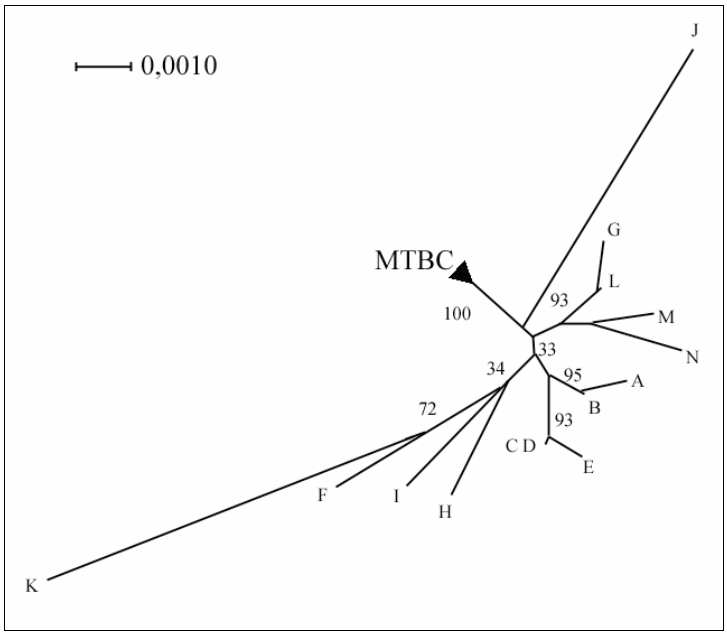


Figure 6 : Relations phylogénétiques (arbre non-enraciné) déduites à partir des 24 séquences concaténées des 12 gènes de ménages. L'échelle représente la distance p. Le triangle noir regroupe les membres du MTBC dont les distances maximales sont égales à 0,1%. Les valeurs de bootstrap obtenues après 1000 itérations sont indiquées sur les branches.

Du fait de l'incongruence observée entre les gènes individuels, nous avons cherché à visualiser l'existence de conflit (relations phylogénétiques incompatibles) entre sites nucléotidiques sur la séquence concaténée en utilisant un arbre en réseau (split decomposition analysis). Selon cette méthode, les parallélépipèdes qui apparaissent (qui peuvent être conçus comme des chemins alternatifs) traduisent le conflit entre sites. Le réseau obtenu montre la présence d'un réseau central de parallélépipèdes très étendu, ce qui traduit l'incompatibilité des sites, effet attendu en cas de recombinaison entre gènes. Les deux groupes identifiés par la méthode Neighbor-Joining sont reconnaissables. Les membres du MTBC forment un complexe clonal supporté par une valeur de bootstrap de 100 % avec des distances phylogénétiques faibles au sein du complexe. L'ordre des embranchements au sein du MTBC n'est pas en contradiction avec les précédentes études (Figure 1). En particulier, la souche TbD1+ considérée comme ancestrale (car ayant encore la région de différence 1) apparaît comme étant la plus proche des souches lisses dans la branche MTBC. Comme avec les arbres NJ, on peut remarquer la distance importante qui sépare les génotypes K et J des autres souches.

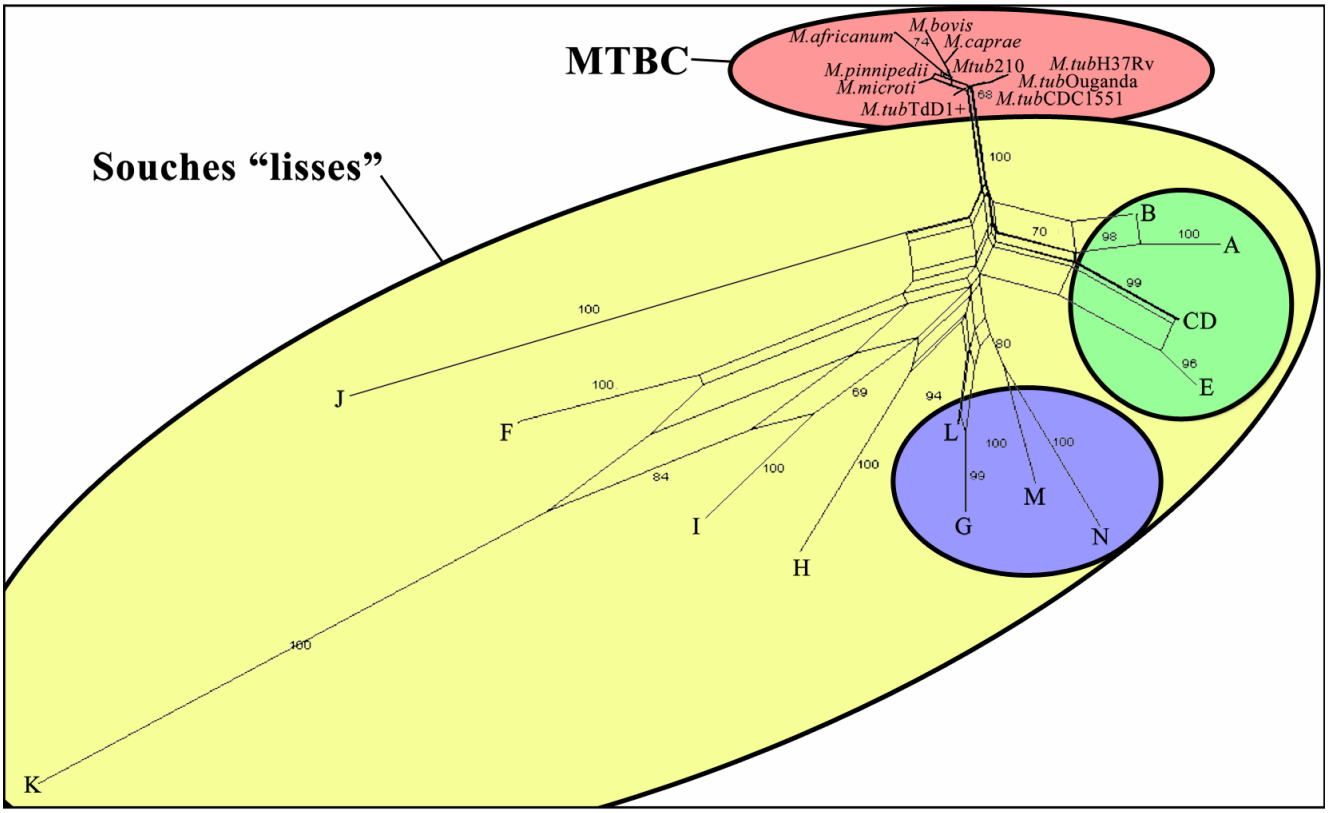


Figure 7 : Split tree réalisé à partir des 24 séquences concaténées des 12 gènes de ménages. Seules les valeurs de bootstrap supérieurs à 69% ont été indiquées.

E. MLST

En cas de présence de recombinaison, il est difficile d’établir les relations de parenté entre souches en utilisant des méthodes basées sur les séquences. En effet, si une séquence très divergente est importée dans une souche, la prise en compte des nombreux nucléotides divergents importés éloignera de manière exagérée, dans les analyses phylogénétiques, cette souche de son génotype ancestral. Pour remédier à cette distorsion causée par la recombinaison homologue, il est préférable de coder les séquences par des numéros d’allèles, et de baser les relations phylogénétiques sur les profils alléliques.

La méthode Minimmum Spanning Tree (réalisée avec le logiciel BioNumerics) trouve le chemin de longueur (somme des différences alléliques) minimale reliant l’ensemble des génotypes entre eux (Figure 8). Dans l’analyse des profils alléliques, un groupe est défini comme un ensemble de souches partageant au moins N allèles communs avec au moins un autre membre du groupe. Dans le cas général où on utilise 7 gènes, on considère qu’un groupe formé avec N = 6 (cas le plus stringent) correspond à un complexe clonal, c’est-à-dire à un ensemble de souches descendant d’une bactérie ancestrale unique. Avec N < 6, il est moins certain que toutes les souches d’un groupe descendent d’un seul ancêtre commun. Dans notre cas, avec 16 gènes, nous utiliserons N = 9. On obtient ainsi trois groupes.

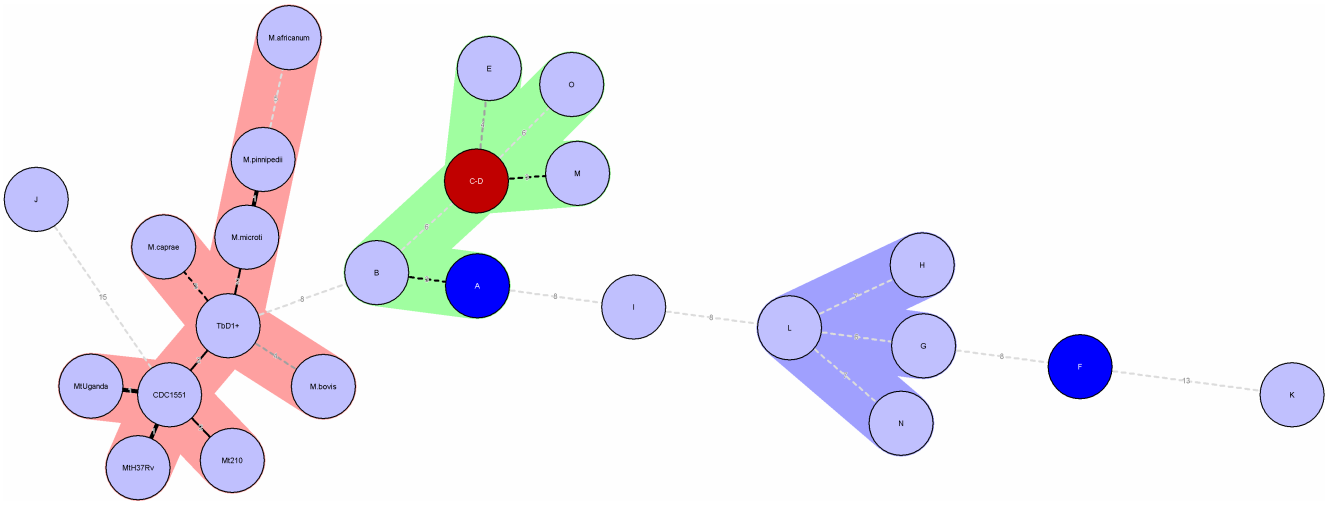


Figure 8 : minimal spanning tree réalisée a partir des profil allèliques de l'ensemble des individus. Chaque disque représente un groupe de souche (en l'occurrence ici les souches ayant le même génotype pour les souches lisses). Ces groupes sont regroupés lorsqu'ils partagent au moins 9 allèles. Le génotype O est représenté ici et partage 11 allèles avec le génotype CD.

Plusieurs observations intéressantes peuvent être formulées. D'une part, les trois groupes révélés correspondent aux trois groupes révélés par les méthodes phylogénétiques, avec deux exceptions. Le groupe M se groupe avec l'ensemble A/B/CD/E ce qui est discordant avec les phylogénies réalisés sur la base des séquences et l'arbre en réseau (Figure 6 et Figure 7) mais s'explique par le nombre de mutations observées (n= 9 pour *gyrA* et *gyrB*) entre les allèles non partagées entre le génotype CD et le génotype M (Tableau 11). De la même manière, le génotype H qui divergeait sur les comparaisons de séquences avec le génotype L, G et N se groupe avec en partageant 9 allèles.

Tableau 11 : Nombre de mutations par gène entre les diffèrents génotypes du groupe A/B/CD/E/M

Génotype comparé	Gènes différents entre les 2 génotypes	Nombre de mutations par gène
A-B	*efp*	1
	hpt	3
	rpoB	2
CD-M	*sodA*	5
	gyrA	9
	gyrB	9
CD-E	*efp*	1
	hsp65	1
	ProS	1
	katG	4

D'autre part, la souche TbD1 est également en position ancestrale au sein du MTBC (au centre du complexe clonal). En comparaison avec les souches lisses, chaque membre de complexe de *M. tuberculosis* se sépare des un des autres par des allèles ne divergeant que d'une seule mutation (Tableau 12).

Finalement, les souches les plus divergentes sur la base des séquences (J et K) le sont également en nombre d'allèles différents. Ces souches paraissent bien les plus divergentes génétiquement, quel que soient le critère utilisé. La souche M2 (groupe O) est apparentée au groupe A/B/CD/E, bien que très éloignée sur la base des séquences concaténées, du fait que les quatre allèles distincts du groupe CD ont des séquences atypiques (voir plus haut).

Tableau 12 : Nombre de mutations par gène entre les différents membres du MTBC

Membre du MTBC comparé		Gènes différents entre les 2 membres du MTBC	Nombre de mutations par gène
M. tuberculosis CDC1551	*M. tuberculosis* Uganda	hsp65	1
M. tuberculosis CDC1551	*M. tuberculosis* H37Rv	gyrA	1
M. tuberculosis CDC1551	*M. tuberculosis* 210	leuS	1
M. tuberculosis CDC1551	*M. tuberculosis* 210	katG	1
M. tuberculosis CDC1551	*M. tuberculosis* TbD1+	gyrB	1
M. tuberculosis CDC1551	*M. tuberculosis* TbD1+	katG	1
M. tuberculosis TbD1+	*M. bovis*	leuS	1
M. tuberculosis TbD1+	*M. bovis*	pncA	1
M. tuberculosis TbD1+	*M. bovis*	gyrB	1
M. tuberculosis TbD1+	*M. caprae*	leuS	1
M. tuberculosis TbD1+	*M. microti*	gyrB	1
M. tuberculosis TbD1+	*M. microti*	adk	1
M. microti	*M. pinnipedii*	gyrB	3
M. pinnipedii	*M. africanum*	adk	1
M. pinnipedii	*M. africanum*	gltX	1
M. pinnipedii	*M. africanum*	glyS	1
M. pinnipedii	*M. africanum*	hsp65	1
M. pinnipedii	*M. africanum*	rpoB	2

Au cours du temps, la diversification des groupes (ou complexes clonaux) peut se faire par recombinaison homologue ou par mutation ponctuelle. Dans ce dernier cas, on observera en général une seule différence nucléotidique entre l'allèle dérivé et ancestral. Dans le premier cas, plus de nucléotides peuvent différer, en fonction de la souche donatrice. Dans nos données, les différences nucléotidiques entre allèles différents pour les souches ne différant pas de plus de trois allèles sur les 16 gènes varient de 1 à 9. Dans 4 cas, on a une seule différence (mutation probable) ; dans 6 cas, on a plus d'une différence (recombinaison probable). Ces données (Tableau 11) montrent que les deux phénomènes semblent avoir une fréquence relativement similaire chez *M. prototuberculosis*. En revanche, 32 nucléotides ont varié par recombinaison, contre 4 seulement par mutation. L'impact de la recombinaison sur la divergence entre séquences est donc approximativement 8 fois plus fort que celui de la mutation. En comparaison, ces rapports recombinaison/mutation sont de 10:1 et 50:1 pour *Streptococcus pneumoniae*, et de 4:1 et 80:1 pour le *Neisseria meningitidis* (Feil et al 2000) [7]. Chez *Staphylococcus aureus*, il est 15 fois plus probable de changer un allèle par mutation que par recombinaison. Chez *E. coli*, il est 10 à 50 fois plus probable de changer un allèle par recombinaison que par mutation. *M. prototuberculosis* semble donc avoir un taux de recombinaison homologue

légèrement inférieur à celui de *N. meningitidis*, mais l'impact sur la diversification des séquences est nettement plus faible, similaire à celui trouvé chez *E. coli*.

IV. Conclusions

Le taux de divergence des séquences au sein d'une espèce bactérienne est très variable d'une espèce à l'autre. Certaines espèces sont très monomorphes, par exemple, *Yersinia pestis* (Achtman et Al 2004) ou *Bordetella pertussis*, l'agent de la coqueluche (Diavatopoulos et al 2005) [1] [5]. Ces espèces ont été élevées au rang taxonomique d'espèces principalement sur le critère de leur importance clinique. D'un point de vue génétique, toutes les souches de ces espèces sont très proches et appartiennent à un même clone dérivant d'un ancêtre commun récent, mais font partie d'un ensemble génétique plus large, dont ce clone est issu. Le MTBC est un autre exemple de groupes monomorphe ou clone, mais son espèce progénitrice, *M. prototuberculosis*, n'a été découverte que récemment. Cette conclusion récente est essentiellement basée sur l'analyse de 6 gènes de ménage, en plus de marqueurs génomiques qui concourent à la même conclusion. Cette hypothèse a été critiquée très récemment (Smith NH. PLoS Pathogens sous presse), en partie à cause du faible nombre de gènes utilisés [18].

L'objectif premier de ce travail était de défendre cette conclusion en utilisant un plus grand nombre de gènes. L'objectif secondaire était de mieux évaluer l'importance de la recombinaison chez *M. prototuberculosis*, et d'analyser un plus grand nombre de souches pour préciser la structure génétique et la diversité de cette espèce.

Nous avons séquencé 16 gènes dont 10 gènes nouveaux pour lesquels les conditions de PCR ont dues être mises au point, sur toutes les souches lisses isolées jusqu'à présent (n = 56) et pour comparaison, sur tous les membres du MTBC (n = 10).

Nos résultats montrent clairement l'homogénéité du groupe formé par les souches lisses et le MTBC. Ce groupe est nettement plus homogène que les espèces bactériennes comparatives utilisées. La conclusion apportée par les six premiers gènes est donc totalement confirmée. Du point de vue purement quantitatif, il n'est pas aberrant de considérer ces souches lisses et le MTBC comme appartenant à un même ensemble biologique. Cependant, du fait de l'importance clinique de *M. tuberculosis* et des autres membres du MTBC, il ne serait pas raisonnable de proposer un changement taxonomique pour ces taxons. Nous nous contentons donc de proposer que l'ensemble homogène formé par les souches lisses et le MTBC soit considéré comme une espèce biologique, sans statut dans la nomenclature. En effet, ce concept d'appartenance à une même espèce biologique a d'importantes conséquences pour l'interprétation de la biologie de ces souches.

L'analyse des 10 gènes additionnels a confirmé l'existence de recombinaison génétique mais pas celle de structure mosaïque, qui n'a pas été retrouvée sur aucun des 10 gènes additionnels. De plus nous

avons pour la première fois évaluée l'impact relatif de la recombinaison et de la mutation chez *M. prototuberculosis.* L'analyse de souches additionnelles a montré une diversité supérieure à celle qui a été révélée sur les souches de l'étude précédente, avec deux génotypes très divergents et un génotype atypique (divergent jusqu'à 8% mais pour quatre gènes seulement). Nous avons identifié l'existence de deux groupes de souches lisses qui paraissent contenir des génotypes plus apparentés entre eux. Ces regroupements ont été révélés par deux approches indépendantes (profils et séquences nucléotidiques).

Table des Illustrations

Bibliographie

1. Achtman M, Morelli, et al. (2004) Microevolution and history of the plague bacillus, Yersinia pestis. Proc Natl Acad Sci U S A. 101 : 17837-42.
2. Brosch R., Gordon SV, et al. (2002) A new evolutionary scenario for the Mycobacterium tuberculosis complex. Proc Natl Acad Sci U S A 99 : 3684 – 3689.
3. Bulut Y., Michelsen K. S., et al (2005) Mycobacterium tuberculosis heat shock proteins use diverse Toll-like receptor pathways to activate pro-inflammatory signals. J. Bio. Chem. 280 : 20961-20967
4. Cole S.T., Brosch R., et al. (1997) Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature 393 : 537 – 544.
5. Diavatopoulos DA, Cummings CA, et al. (2005) Bordetella pertussis, the causative agent of whooping cough, evolved from a distinct, human-associated lineage of B. bronchiseptica. PLoS Pathog.1 : e45.
6. Fabre M., Koaeck J.L, et al. (2004) High Genetic Diversity Revealed by Variable-Number Tandem Repeat Genotyping and Analysis of hsp65 Gene Polymorphism in a Large Collection of “Mycobacterium canettii” Strains Indicates that the M. tuberculosis Complex Is a Recently Emerged Clone of “*M. canettii*”. J clin Microbiol 42 : 3248-3255.
7. Feil EJ, Enright MC, Spratt BG (2000) Estimating the relative contributions of mutation and recombination to clonal diversification: a comparison between Neisseria meningitidis and Streptococcus pneumoniae. Res Microbiol. 151:465-9.
8. Fleischmann R. D., Alland D., et al. (2002) Whole-genome comparison of Mycobacterium tuberculosis clinical and laboratory strains. J. Bacteriology 184 : 5479 – 5490.
9. Garnier T., Eiglmeier K., et al. (2003) The complete genome sequence et Mycobacterium bovis. Proc Natl Acad Sci 100 : 7877 – 7882
10. Gil R., Silva F. J., et al (2004) Determination of the core of a minimal bacterial gene set. Micr. Mol Biol. 68 : 518 – 537.
11. Gordon S. V., Heym B. et al. (1999) New insertion sequences and a novel repeated sequences in the genome of Mycobacterium tuberculosis H37Rv. Microbiology 145 : 881 – 892.

12. Gutacker M. M., Smoot J. C., et al (2002) Genome-wide analalysis of synonymous single nucelotide polymorphisms in Mycobacterium tuberculosis complex organism : reslution of genetic relationships among closely related microbial strains. Genetics 162 : 1533 – 1543.

13. Gutiérrez M. C., Brisse S., et al. (2005) Ancient origin and gene mosaicism of the progenitor of Mycobacterium tuberculosis. Plos Pathogens 1 : 55 – 61.

14. Hughes A. L.., Friedman R., et al. (2002) Genowide pattern of synonymous nucleotide substitution in two complete genomes of Mycobacterium tuberculosis. Emerg. Inf. Dis. 8 : 1342 – 1346.

15. Pym A.S., Domenech P., et al. (2001) Regulation of catalase-peroxidase (katG) expression, isonazid sensivity and virulence by furA of Mycobacterium tuberculosis. Mol. Microbiology 40 : 879-889.

16. Saint-Joanis B ;, Souchon H., et al. (1999) Use of-directed mutagenesis to probe structure, function and ionazid activation of the catalse/peroxidase, katG, from Mycobacterium tuberculosis. Biochem. J. 338 : 753-760.

17. Sirakova T.D. , Dubey V. S. et al (2003) The largest open reading frame in Mycobacterium tuberculosis genome is involved in pathogenesis and Dymycoseryl Phthiocerol Synthesis. Inf and Imm. 71 : 3794 - 3801

18. Smith N (2006) A Re-Evaluation of *M. prototuberculosis* . Plos Pathogens 2 : 5 - 7

19. Spratt B. G. (2004) Exploring the concept of clonalty in bacteria. Methods Mol Biol 266 : 323 - 352

20. Sreevatsan S, Pan X, Stockbauer KE, et al. (1997) Restricted structural gene polymorphism in the Mycobacterium tuberculosis complex indicates evolutionarily recent global dissemination. Proc Natl Acad Sci U S A 94: 9869–9874.

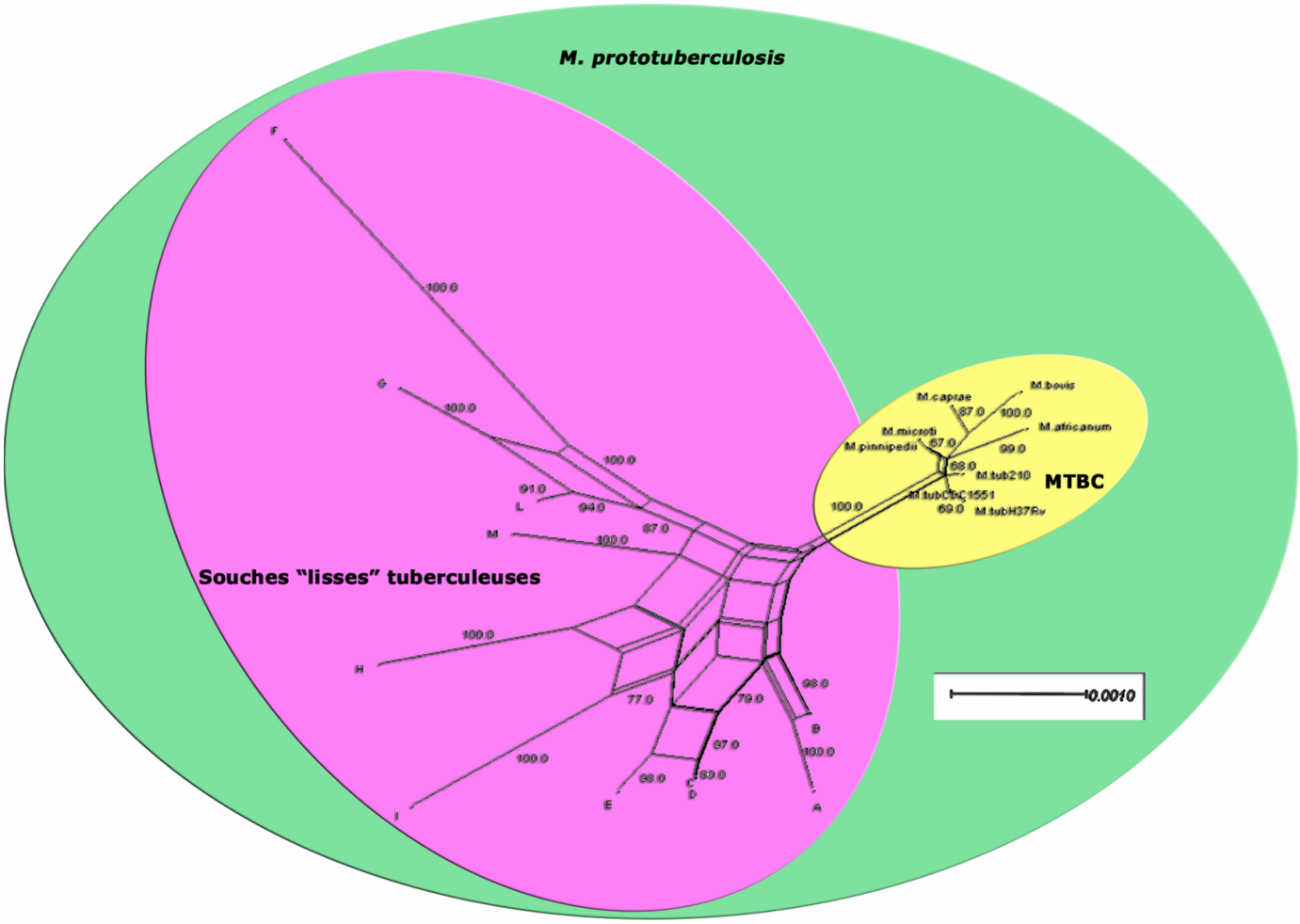
21. Templeton A. R. (2002) Out of Africa again and again. Nature 416 : 45 -51.

22. Van Soolingen D., Hoogenboezem T. et al. (1997) A novel pathogenic taxon of the Mycobacterium tuberculosis complex, Canettii : characterization of an exceptional isolate from Africa. Int. J . Syst. Bact. 47 : 1236 -1245.

23. World Health Organization. (2006) Global tuberculosis control : surveillance, planning, financing. WHO Library cataloguing-in-publication data.

Annexes

Annexe 1 : Split tree réalisé à partir des 18 séquences concaténées des 16 gènes de ménages. Seules les valeurs de bootstrap supérieurs à 69% ont été indiquées. L'échelle représente la distance p non corrigée.



Annexes 2 : représentation des graphique des phylogénies non enracinée des 16 gènes séquencés basées sur les génotypes des souches lisses et les membres du MTBC. L'échelle représente la distance p (non corrigée). Les phylogénies réalisées à partir des 16 et 12 gènes concaténées y figurent également.


hpt
0.01
O
N
M.tubUganda
M.tubTbD1
M.tubH37Rv
M.tubCDC1551
M.tub210
M.pinnipedii
M.microti
M.caprae
M.bovis
M.africanum
M
L
J
H
G
E
D
C
B
I
K
F
A
hsp65
0.04
M.tubTbD1
M.tubH37Rv
M.tubCDC1551
M.tub210
M.pinnipedii
M.microti
M.bovis
M.africanum
M.tubUganda
M.caprae
O
J
M
L
K
I
H
G
F
E
D
C
B
A
katG
0.002
N
M.tubTbD1
M.pinnipedii
M.microti
K
J
L
H
G
F
E
M.tub210
M.caprae
M.bovis
O
M.africanum
M
I
D
C
B
M.tubUganda
M.tubH37Rv
M.tubCDC1551
A
leuS
0.004
M.caprae
H
O
M.tubUganda
M.tubTbD1
M.tubH37Rv
M.tubCDC1551
M.pinnipedii
M.microti
M.africanum
M
I
E
D
C
L
G
J
K
F
M.tub210
B
M.bovis
A
pgi
0.01
M
L
O
M.tubUganda
M.tubTbD1
M.tubH37Rv
M.tubCDC1551
M.tub210
M.pinnipedii
M.microti
M.caprae
M.bovis
M.africanum
J
H
K
F
I
G
N
E
D
C
B
A
pncA
0.007
M.tubUganda
M.tubTbD1
M.tubH37Rv
M.tubCDC1551
M.tub210
M.pinnipedii
M.microti
M.caprae
M.bovis
M.africanum
M
O
L
I
H
E
D
C
N
J
F
K
G
B
A
proS
0.005
N
M.tubH37Rv
M.tubCDC1551
M.tub210
M.pinnipedii
M.microti
M.caprae
M.bovis
M.africanum
L
I
H
G
F
E
M
K
J
D
C
B
A
recA
0.004
M.tubTbD1
M.tubH37Rv
M.tubCDC1551
M.pinnipedii
M.microti
M.tubUganda
M.tub210
M.caprae
M.bovis
M.africanum
G
F
O
M
E
D
C
N
L
I
H
B
A
rpoB
0.04
M.tubUganda
M.tubTbD1
M.tubH37Rv
M.tubCDC1551
M.tub210
M.pinnipedii
M.microti
M.africanum
M.caprae
M.bovis
M
K
O
J
F
E
D
C
B
N
L
I
H
G
A