La mise en place de bases de données des génotypes circulants du complexe Mycobacterium tuberculosis et des outils web permettant de mieux surveiller, comprendre et contrôler l'épidémie de la tuberculose dans le monde

David Couvin, Nalin Rastogi

Réseaux

La mise en place de bases de données des génotypes circulants du complexe *Mycobacterium tuberculosis* et des outils web permettant de mieux surveiller, comprendre et contrôler l'épidémie de la tuberculose dans le monde

David Couvin et Nalin Rastogi (nrastogi@pasteur-guadeloupe.fr).
Laboratoire supranational de référence pour la TB de l'OMS, institut Pasteur de la Guadeloupe, Abymes, Guadeloupe, France

Mots-clés : *Mycobacterium tuberculosis*, tuberculose, génotypage, bases de données, spoligotypage, MIRU-VNTR, épidémiologie, phylogénie, résistance aux médicaments.

Résumé

Dans cet article, nous présenterons rapidement plusieurs bases de données de génotypage du complexe *Mycobacterium tuberculosis* (MTBC), élaborées au cours des quinze dernières années à l'Institut Pasteur de la Guadeloupe (IPG), dans le cadre d'une grande initiative concertée visant à lutter contre l'épidémie mondiale de la tuberculose. De la toute première version au format Excel en 1999 (SpolDB1 ; n = 610 isolats cliniques) à la quatrième au format MySQL en 2006 (SpolDB4 ; n = 39 295 isolats cliniques), ces bases de données ont brossé un premier tableau de la phylogéographie des lignées génotypiques de MTBC en circulation, en utilisant le typage oligonucléotidique des espaceurs (spoligotypage) qui permet d'étudier le polymorphisme du locus DR (*Direct Repeat*). Les deux dernières versions multimarqueurs sous MySQL constituent respectivement la 5e version de SITVITWEB, publiée en 2012 (n = 62 582 isolats cliniques), qui repose sur le spoligotypage et le typage MIRU-VNTR (*Mycobacterial Interspersed Repetitive Units/Variable Number of Tandem Repeats*) 12 loci, et la 6e version, dénommée SITVIT2, qui sera publiée en 2014 (n = 111 635 isolats cliniques) et qui contient des données de spoligotypage et de MIRU-VNTR 12, 15 ou 24 loci. Sur ces dernières versions, une interface en ligne permet à l'utilisateur de rechercher des souches dans la base de données au moyen de critères tels que l'année, le pays d'isolement, le pays d'origine ou le nom du chercheur. Cette interface permet en outre d'effectuer des recherches mixtes dans SITVIT2, permettant d'obtenir les données de génotypage de certaines souches, conjointement avec leur répartition géographique, ainsi que les données disponibles sur la résistance aux antibiotiques ou les caractéristiques démographiques et épidémiologiques. Notre initiative de recherche a ainsi pour but d'améliorer la caractérisation phylogénétique détaillée des lignées du MTBC, ainsi que l'épidémiologie des clones en circulation, afin d'élaborer une cartographie géographique factuelle des isolats cliniques prédominants pour les bacilles tuberculeux principalement impliqués dans la maladie, à l'échelle nationale et régionale. La superposition ultérieure de ces cartes avec des données sociopolitiques, économiques et démographiques obtenues auprès de Systèmes d'information géographique (SIG) dressera un portrait précis des disparités actuelles par sous-région, selon la classification des Nations Unies. Il est important de comprendre le détail de ces disparités et faiblesses pour que les décideurs politiques et les autorités de santé publique puissent prendre les mesures qui s'imposent pour mieux surveiller, comprendre et contrôler l'épidémie mondiale de la tuberculose.

Introduction

Près de vingt ans après que la tuberculose (TB) a été déclarée « urgence de santé publique mondiale » par l'Organisation mondiale de la santé (OMS), et malgré les nets progrès réalisés pour atteindre les objectifs internationaux de 2015, fixés dans le cadre des objectifs du Millénaire pour le développement, la TB demeure la deuxième maladie infectieuse la plus meurtrière au monde après le VIH/SIDA (rapport de l'OMS, WHO 2013). Selon les estimations, la TB a été responsable de 8,6 millions de nouveaux cas et de 1,3 million de décès (dont 320 000 chez des patients co-infectés par la TB et le VIH) en 2012. Un examen attentif du rapport de l'OMS révèle que la TB demeure un problème économique et sanitaire grave, non seulement dans les pays en développement, mais aussi au sein des nations développées, en raison des co-infections TB/VIH et de l'émergence de souches multirésistantes aux médicaments (*Multi-Drug Resistant*, MDR) et, plus récemment, de souches extrêmement résistantes aux médicaments (*eXtremely Drug Resistant*, XDR), complexifiant la gestion de la maladie et augmentant considérablement la mortalité par TB chez les patients immunodéprimés.

De plus, l'intensification des déplacements et des migrations de population à laquelle on assiste depuis plusieurs décennies, que ce soit pour des motifs touristiques ou professionnels, soulève un nouveau problème dans les pays où la TB était en déclin : la modification des scénarios socioépidémiologiques en raison d'une immigration massive en provenance de pays où la TB est fortement endémique (García de Viedma *et al.,* 2011). Parmi les questions qui se posent immanquablement, on retiendra notamment : la comparaison entre le rôle de la transmission récente et celui de la réactivation/importation dans les cas de TB d'origine étrangère ; les répercussions d'une éventuelle importation de souches de *M. tuberculosis* non identifiées à ce jour ; et la transmission croisée entre cas de différentes nationalités. Il est donc important de comprendre comment se transmettent les bacilles tuberculeux, de savoir quels sont les clones impliqués dans les épidémies et/ou les cas résistant à l'action des médicaments, d'identifier de nouveaux clones susceptibles de voir le jour dans un milieu donné ou d'être en voie d'extinction, d'identifier les sous-populations les plus à risque de contracter l'infection et les facteurs de risque associés, et de pouvoir interpréter ces résultats à la lumière de l'évolution du complexe *Mycobacterium tuberculosis* (MTBC). Il est évident qu'il serait difficile, aujourd'hui, de répondre à ces questions sans l'aide de l'épidémiologie moléculaire.

On considérait, il y a encore peu de temps, que toutes les souches de MTBC adaptées à l'Homme étaient pratiquement identiques ; la question de la variabilité génétique individuelle

Réseaux

entre espèces du MTBC n'a donc pas retenu une grande attention et la plupart des travaux réalisés en la matière se sont focalisés sur des organismes précis. L'arrivée des techniques moléculaires et leur généralisation dans les études des populations de bacilles tuberculeux ont fait évoluer les idées et les technologies. Bien que le MTBC constitue un groupe remarquablement homogène sur le plan génétique, avec des signes tangibles d'évolution clonale, des études récentes ont démontré que la diversité génétique entre les différents clones était nettement plus importante qu'on ne le pensait, ce qui pourrait avoir un impact sur leurs propriétés pathobiologiques. Une étude de référence menée sur une collection mondiale de souches de MTBC, à partir des données de séquence d'ADN de sept paires de mégabases, a ainsi révélé une diversité génétique importante (Hershberg *et al.,* 2008). Les auteurs de cette étude pensent qu'une grande partie de cette diversité génétique provient d'une dérive génétique qui pourrait être liée à des évènements démographiques et migratoires humains, avec des conséquences fonctionnelles comme l'émergence et la propagation d'une TB résistante aux médicaments.

Bien que la TB soit présente dans le monde entier, son incidence est plus élevée dans certains pays ou certaines régions que dans d'autres : l'Afrique subsaharienne, l'Asie du Sud et du Sud-Est, l'Amérique latine, mais aussi l'Europe de l'Est et la Russie, ainsi que les Antilles (notamment Haïti et la République dominicaine), par exemple. En effet, la TB constitue toujours un problème sanitaire grave dans de nombreux pays des Antilles et d'Amérique latine ; par exemple, l'incidence de la TB à Haïti, qui s'élevait à 330 cas/100 000 habitants dans les années 90, est toujours aussi élevée en 2012, avec 213 cas/100 000 habitants (rapport de l'OMS, 2013). Il va sans dire que le diagnostic et la confirmation bactériologique précoces de la TB ainsi que la détermination de la résistance aux médicaments constituent des étapes clés dans la lutte contre l'épidémie de TB, et que les laboratoires de référence constituent des structures de premier plan en matière de diagnostic et de lutte antituberculeuse à l'échelle mondiale. Dans un tel contexte, l'Institut Pasteur a pris l'initiative de créer le premier laboratoire de référence pour la TB dans la région des Antilles en 1993. Situé à l'Institut Pasteur de la Guadeloupe (IPG), ce laboratoire était initialement destiné à travailler pour les Antilles ; cependant, bien conscients de la portée mondiale de la TB, nous avons dès le départ adopté une approche fondée sur l'utilisation concomitante de techniques bactériologiques et moléculaires.

Au cours des 20 dernières années, nous avons mis au point une approche complète qui intègre le dépistage moléculaire systématique du MTBC et d'autres espèces mycobactériennes, la surveillance de la résistance aux antibiotiques, et l'élaboration de méthodes de diagnostic rapide et de diverses techniques moléculaires utiles pour l'épidémiologique et les études des populations de bacilles tuberculeux aux niveaux local, régional et mondial. Pour la surveillance de la TB à l'échelle mondiale, nous avons créé, depuis 1999, une série de bases de données de génotypage qui regroupent non seulement nos propres données, mais aussi celles recueilles auprès de différents laboratoires participants implantés dans le monde entier. Cette publication rappellera brièvement les étapes suivies pour constituer ces bases de données et les outils en ligne mis au point pour permettre de créer des cartes de répartition géographique mondiale montrant la fréquence des génotypes de TB dans le monde à différentes échelles géographiques. Nous citerons également quelques études, publiées ou en cours, qui utilisent ces données et nous présenterons les perspectives qui se dessinent aujourd'hui.

TB : origine, propagation et co-adaptation avec les hôtes

Depuis l'isolement et la caractérisation d'un ancien ADN de *M. tuberculosis* datant de 17 000 ans avant J. C. chez une espèce de bison disparue, révélant la présence de la TB en Amérique vers la fin du pléistocène (Rothschild *et al.,* 2001), on sait que la TB est une maladie très ancienne. L'étude d'ADN anciens prélevés sur des restes humains a permis de détecter la présence de la TB sur des momies égyptiennes, avec la caractérisation de *M. tuberculosis* et *M. africanum* (Zink *et al.,* 2003). On a alors pensé, par analogie avec d'autres épidémies liées à l'expansion démographique (« crowd diseases"), que l'origine de la TB humaine était liée à la transition démographique néolithique (TDN), qui a commencé il y a environ 11 000 ans. En effet, le développement de la domestication des animaux a accru le risque de transmission zoonotique de nouveaux agents pathogènes à l'Homme, tandis que les innovations agricoles ont favorisé l'accroissement des densités de population, contribuant à entretenir le cycle infectieux (Wolfe *et al.,* 2007). On ne savait cependant pas avec certitude (a) si la TB humaine descendait d'une mycobactérie de ruminant qui avait récemment infecté l'Homme par l'intermédiaire des animaux domestiques ou d'une ancienne mycobactérie humaine qui avait infecté les ruminants sauvages et domestiques ; et (b) si la TB était apparue indépendamment dans les deux hémisphères ou si elle avait été amenée aux Amériques par les Européens. Toutefois, la TB présentait aussi un profil de chronicité, de latence et de réactivation caractéristique d'une maladie pré-TDN (Barry *et al.,* 2009).

C'est grâce à de nouvelles recherches qui ont montré que la TB est probablement aussi vieille que l'humanité (Comas *et al.,* 2013) que l'on a obtenu la réponse à ces questions. En étudiant les différentes variations génétiques au sein du MTBC, des chercheurs sont parvenus à démontrer que la TB s'est certainement propagée dans le monde avec les premiers Hommes modernes ayant quitté l'Afrique. Cette étude a consisté à analyser l'ensemble des génomes d'une collection de 259 souches de MTBC contemporaines provenant du monde entier, puis à comparer la diversité phylogénétique du MTBC à la diversité humaine, déduite de données sur le génome mitochondrial. Les résultats obtenus ont révélé que le MTBC est apparu il y a environ 70 000 ans, qu'il a accompagné les migrations des Hommes anatomiquement modernes hors d'Afrique et qu'il s'est propagé en raison de l'accroissement des densités de population humaine pendant la période néolithique (Comas *et al.,* 2013). Cette origine ancienne de la TB et le fait que la phylogénie génomique du MTBC soit calquée sur celle des génomes mitochondriaux humains ont par ailleurs révélé que la forme pulmonaire de la TB ne s'est pas propagée à l'Homme par l'intermédiaire des animaux domestiques, puisque l'élevage est apparu beaucoup plus tard. Cette étude de référence a également révélé des similitudes frappantes au niveau de l'évolution de l'Homme et du MTBC, suggérant non seulement que l'évolution du MTBC a suivi celle de l'Homme, mais aussi que la diversité du MTBC a directement tiré profit des explosions démographiques humaines. Étant donné que *M. tuberculosis* est un pathogène humain strict, sans réservoir animal ou environnemental connu, il est très probable que les changements démographiques et l'accroissement des densités

Réseaux

de population humaine qui s'opèrent au fil du temps affectent son évolution, comme dans le cas d'un modèle de maladie de groupe « crowd diseases ». En même temps, sa latence et sa chronicité peuvent lui permettre de s'adapter à des densités d'hôtes plus faibles, de survivre, puis de frapper à nouveau lorsque les conditions sont favorables à une infection massive.

Techniques de typage pour l'épidémiologie moléculaire de la TB

Le MTBC est un groupe d'organismes très varié sur le plan écologique ; il comprend *M. tuberculosis, M. africanum* et *M. canettii* (pathogènes humains stricts), M. microti (pathogène des rongeurs) et *M. bovis* (pathogène des bovins qui présente cependant un large spectre d'hôtes), ainsi que *M. pinnipedii* (phoques), *M. caprae* (chèvres), *M. mungi* (mangoustes rayées) et les bacilles de l'oryx (rebaptisé *M. orygis*), du rat des rochers et du chimpanzé, agents étiologiques de la TB chez les espèces animales qui ont donné leur nom (bacilles de « dassie » et « chimpanzee" respectivement). Cependant, le MTBC est un groupe très homogène sur le plan génétique et plusieurs de ses membres partagent, en moyenne, plus de 99,7 % d'identité nucléotidique (Kato-Maeda *et al.,* 2001). Malgré cette remarquable homogénéité génétique, les 20 dernières années ont vu naître de nouvelles techniques moléculaires, largement utilisées aujourd'hui dans les études de génotypage des populations, qui permettent de caractériser les isolats de TB avec précision et de déduire les différentes lignées phylogénétiques qui y sont associées. Des revues détaillées sont disponibles sur l'évolution moléculaire du MTBC (Rastogi et Sola, 2007), les techniques de typage moléculaire actuelles (Jagielski *et al.,* 2014), ainsi que des stratégies et innovations dans le vaste domaine de l'épidémiologie moléculaire de la TB (2) (García de Viedma *et al.,* 2011) ; les lecteurs intéressés sont invités à consulter ces références pour approfondir ces questions.

Bien que la méthode IS6110-RFLP ait longtemps été considérée comme la technique de référence pour le typage de *M. tuberculosis* en raison de sa reproductibilité et de son pouvoir discriminant pour l'épidémiologie moléculaire de la TB (van Embden *et al.,* 1993), cette méthode laborieuse nécessitait de grandes quantités d'ADN et se caractérisait par l'absence de pouvoir discriminant pour le typage des isolats ayant un faible nombre de copies d'IS6110, par exemple dans le sud de l'Inde (Radhakrishnan *et al.,* 2001). De surcroît, l'horloge moléculaire rapide de ce marqueur pour les études d'évolution (de Boer *et al.,* 1999), la complexité des forces qui régissent sa transposition et le risque de convergence génétique (Fang *et al.,* 2001), ainsi que la difficulté à élaborer de grandes bases de données RFLP et la nécessité de disposer de logiciels perfectionnés pour l'analyse des données (Heersma *et al.,* 1998 ; Salamon *et al.,* 1998) ont mis en doute l'intérêt de son utilisation en génétique de l'évolution. Pour les raisons exposées plus haut, la technique IS6110-RFLP a largement fait place, au cours des 10 dernières années, à de nouvelles stratégies de typage par PCR comme le spoligotypage (Kamerbeek *et al.,* 1997) et la technique MIRU-VNTR (*Mycobacterial Interspersed Repetitive Units/Variable Number of Tandem Repeats*) (Supply *et al.,* 2001 ; 2006), jetant ainsi les bases d'un génotypage de *M. tuberculosis* à haut débit et à grande échelle.

S'appuyant sur le polymorphisme du locus DR (*Direct Repeat*), le spoligotypage (de l'anglais SPacer OLIGOnucleotide TYPING, typage oligonucléotidique des espaceurs) fait actuellement partie des techniques de PCR les plus utilisées pour l'épidémiologie moléculaire et l'étude phylogéographique du MTBC. D'abord identifié dans la souche vaccinale M. bovis BCG (Hermans *et al.,* 1991), le locus DR appartient à la famille d'ADN répétitifs CRISPR (*Clustered Regularly Interspaced Short Palindromic Repeats*) et contient plusieurs répétitions parfaites de 36 pb (3 tours d'hélice) identiques, séparées par des espaceurs de 34 à 41 pb uniques. Avec leurs espaceurs non répétitifs, les motifs DR forment plusieurs répétitions directes variables (DVR, Direct Variant Repeats) qui présentent un important polymorphisme entre les isolats cliniques de *M. tuberculosis*. Une étude a précédemment démontré que le profil binaire à 43 caractères obtenu par cette technique véhiculait des informations phylogénétiques déterminantes (Sola *et al.,* 2001). Aujourd'hui, les profils de spoligotypage sont généralement désignés par leur code octal, format descriptif faisant l'objet d'une convention internationale (Dale *et al.,* 2001). En raison de sa simplicité, de son format de résultats binaire et de sa grande reproductibilité, le spoligotypage est très utilisé pour l'épidémiologie moléculaire du MTBC, en tant que technique sur macroréseaux (*macroarrays*), pour analyser la présence ou l'absence de 43 espaceurs prédéfinis (sur 104 espaceurs présents). En effet, la recherche « spoligotyping OR spoligotype » dans PubMed donne 924 articles publiés (interrogation effectuée le 18 mars 2014).

L'une des limites du spoligotypage étant sa tendance à surestimer le regroupement en grappes du MTBC, il a rapidement été proposé de compléter cette technique par un typage reposant sur des minisatellites (MIRU-VNTR), dans le cadre d'une stratégie « à double PCR », en association avec les études épidémiologiques classiques (Sola *et al.,* 2003). Les minisatellites MIRU-VNTR constituent un ensemble de marqueurs multi-loci dans la mesure où ils représentent des marqueurs indépendants d'un même type ; ils sont utilisés aux formats 12 loci classique, 15 loci discriminant ou 24 loci complet (Supply *et al.,* 2001 ; 2006). Cependant, même les MIRU 24 loci ne présentent pas un pouvoir de résolution satisfaisant pour discriminer avec précision des souches de génotype Beijing étroitement apparentées, si bien qu'il a récemment été proposé d'utiliser un autre ensemble de MIRU-VNTR consensus hypervariables 4 loci pour le sous-typage des grappes et complexes clonaux du génotype Beijing (Allix-Béguec *et al.,* 2014).

Techniques de typage moléculaire pour la phylogénie de la TB

Outre leur utilisation en épidémiologie (García de Viedma *et al.,* 2011), les techniques de typage moléculaire sont également utilisées dans les études d'évolution (Rastogi et Sola, 2007). Elles comprennent alors essentiellement deux ensembles de marqueurs : (a) ceux cités plus haut, qui sont largement utilisés dans les études épidémiologiques, mais qui fournissent également des informations phylogénétiques concomitantes – IS6110-RFLP, spoligotypage et MIRU-VNTR (voir ci-après) ; et (b) un ensemble de marqueurs comprenant les polymorphismes de longues séquences (LSP, *Large Sequence Polymorphisms*)/régions de différence (RD) et les polymorphismes mononucléotidiques (SNP, *Single Nucleotide Polymorphisms*), particulièrement utiles pour les études de phylogénie et d'évolution.

L'une des toutes premières études en la matière a eu recours à l'hybridation génomique soustractive pour identifier trois

Réseaux

régions génomiques distinctes, entre des souches virulentes de *M. bovis* et *M. tuberculosis* et la souche avirulente *M. bovis* BCG, respectivement dénommées RD1, RD2 et RD3 (Mahairas *et al.,* 1996). Dans une autre étude, les chercheurs sont parvenus à distinguer trois groupes génétiques de *M. tuberculosis* sur la base de deux polymorphismes apparaissant avec une fréquence élevée dans les gènes codant la catalase-peroxydase et la sous-unité A de la gyrase, aboutissant à une classification en trois groupes génétiques principaux (PGG, *Principal Genetic Groups*), les bactéries du groupe 1 étant ancestrales de celles des groupes 2 et 3 (Sreevatsan *et al.,* 1997). Presque aussitôt, des matrices de chromosomes artificiels bactériens (BAC, *Bacterial Artificial Chromosome*) de la souche H37Rv digérés par une enzyme de restriction ont été utilisées pour révéler la présence de 10 RD entre *M. tuberculosis* et *M. bovis* (RD1 à RD10), dont 7 (RD4 à RD10) étaient délétées chez *M. bovis* (Gordon *et al.,* 1999). Dans une présentation de référence, Brosch *et al.* (2002) ont analysé la répartition de 20 régions variables résultant d'évènements d'insertion-délétion dans le génome des bacilles tuberculeux d'une collection de souches appartenant à toutes les sous-espèces du MTBC, démontrant que, sur la base de la présence ou de l'absence d'une délétion 1 spécifique de *M. tuberculosis* (TbD1, séquence de 2 kb), *M. tuberculosis* pouvait être subdivisé en souches TbD1-positives « anciennes » et en souches TbD1-négatives « modernes » (Brosch *et al.,* 2002). Selon ce nouveau scénario de l'évolution du MTBC, la délétion de RD9 identifie une lignée, représentée par *M. africanum*, *M. microti* et *M. bovis*, qui a divergé de l'ancêtre des souches de *M. tuberculosis* actuelles avant l'apparition de TbD1, découverte qui contredit les hypothèses antérieures selon lesquelles *M. tuberculosis* a évolué à partir d'un ancêtre de *M. bovis* (Brosch *et al.,* 2002). Étant donné qu'aucune de ces régions n'est manquante chez *M. canettii* et d'autres souches ancestrales du MTBC, l'on présume que ce sont des descendants directs des bacilles tuberculeux qui existaient avant la séparation entre la lignée « *M. africanum* – *M. bovis* » et la lignée « *M. tuberculosis* ».

En utilisant une collection mondiale de MTBC et 212 SNP, Filliol *et al.* (2006) ont identifié six groupes de grappes SNP (SCG, *SNP Cluster Groups*) phylogénétiquement différents et à embranchement profond, ainsi que cinq sous-groupes. Ces SCG sont fortement corrélés avec l'origine géographique des échantillons de *M. tuberculosis* et le lieu de naissance des hôtes humains. Les auteurs de l'étude ont proposé un algorithme capable d'identifier deux ensembles minimaux composés de 45 ou de 6 SNP, sur 212 SNP testés, qui peuvent servir à cribler des collections mondiales de MTBC pour étudier l'évolution, la différenciation des souches ou les différences biologiques entre souches. Dans une autre étude, Gutacker *et al.,* (2006) se sont intéressés aux relations génétiques du MTBC en analysant 36 sSNP au sein d'une grande collection de souches isolées de patients inclus dans quatre études des populations aux États-Unis et en Europe, et ils ont affecté cette collection à 1 de 9 grandes grappes génétiques. Une classification similaire a été révélée par analyse d'un panel élargi à d'autres SNPs. Étant donné que les profils de classification des lignées phylogénétiques typées par SNP ont été associés de manière non aléatoire à des profils IS6110, des spoligotypes et des profils MIRU-VNTR, les auteurs pensent que la population du MTBC présente une structure fortement clonale.

Parallèlement, en utilisant des microréseaux d'ADN (*microarrays*) pour identifier avec précision les LSP, des chercheurs ont observé une corrélation stable entre les souches de MTBC et leurs populations d'hôtes humains (Hirsh *et al.,* 2004) ; les analyses phylogénétiques ont non seulement indiqué que les transferts horizontaux de gènes étaient rares au sein du MTBC, mais aussi que les corrélations entre les populations d'hôtes et d'agents pathogènes étaient stables, même dans un milieu urbain cosmopolite (comme San Francisco) et qu'elles dépendaient largement de la composition de la population locale d'immigrés. Les auteurs de cette étude ont conclu que *M. tuberculosis* est organisé en plusieurs grandes populations génétiquement différenciées qui sont à leur tour corrélées directement et de manière stable avec des populations d'hôtes délimitées selon leur lieu d'origine. Une publication postérieure de la même équipe de chercheurs a confirmé cette compatibilité variable entre hôtes et agents pathogènes, la structure de la population mondiale de *M. tuberculosis* étant définie par six lignées phylogéographiques typées par RD/LSP, dont chacune est associée à des populations humaines sympatriques spécifiques : la lignée indo-océanique, la lignée est-asiatique, la lignée est-africaine/indienne, la lignée euro-américaine et deux lignées ouest-africaines (Gagneux *et al.,* 2006).

Tableau 1. Comparaison de la nomenclature des lignées de *M. tuberculosis* fondée sur le spoligotypage avec celles fondées sur les PGG, les groupes de grappes SNP (SCG), le typage SNP et le typage LSP.

Spoligotyping-based (Filliol 2003)	PGG	SCG (Filliol 2006)	SNP-based (Gutacker 2006)	LSP (Gagneux 2006)
East-African-Indian (EAI)	PGG1	SCG 1	sSNP-I	Indo-Oceanic
Beijing	PGG1	SCG 2	sSNP-II	East-Asian
Central-Asian (CAS)	PGG1	SCG 3a	sSNP-IIA	East-African-Indian
Haarlem	PGG2	SCG 3b	sSNP-III	Euro-American
X1	PGG2	SCG 3c	sSNP-IV	Euro-American
X1,X2,X3	PGG2	SCG 4	sSNP-V	Euro-American
LAM	PGG2	SCG 5	sSNP-VI	Euro-American
T (Miscellaneous)	PGG2-3	SCG 6	sSNP-VII sSNP-VIII	Euro-American
Bovis	PGG1	SCG 7	(MTBC)	(MTBC)
M. africanum	PGG1	NA	NA	West-African 1
M. africanum	PGG1	NA	NA	West-African 2

Le **tableau 1** récapitule les correspondances existant entre différentes nomenclatures des lignées génotypiques. Il convient de souligner qu'il faut retenir le nom du marqueur utilisé lorsque l'on désigne une lignée, en particulier la lignée « East-African Indian » (EAI) qui fait référence à deux groupes de *M. tuberculosis* totalement différents selon qu'il s'agit du spoligotypage ou du typage LSP. Une observation intéressante est la bonne congruence qui existe entre le spoligotypage et le typage SNP (Filliol *et al.,* 2006), les spoligotypes EAI et « Beijing » coïncidant respectivement avec les groupes SCG 1 et SCG 2 ; les spoligotypes X et « Central Asian » (CAS) sont également associés à un SCG et/ou sous-groupe donnée. La corrélation avec les SCG n'est pas aussi franche pour les autres lignées. De plus, les différentes lignées identifiées par spoligotypage coïncident bien avec les anciens PGG, si bien

Réseaux

que l'on peut provisoirement classer les souches du MTBC en un groupe TbD1+/PGG1 ancestral (sous-ensemble 1 : *M. africanum* et EAI), un groupe TbD1–/PGG1 moderne (sous-ensemble 2 : « Beijing » et CAS) et un groupe TbD1–/PGG2/3 d'évolution récente (sous-ensemble 3 : « Haarlem », X, S, T et « Latin American and Mediterranean » ou LAM). Néanmoins, les inférences épidémiologiques et phylogénétiques ne sont pas toujours faciles à faire en raison du manque de connaissance des mécanismes mutationnels qui donnent lieu au polymorphisme de ces cibles génomiques. Des études récentes ont démontré que des souches de MTBC non phylogénétiquement apparentées pouvaient parfois présenter le même profil de spoligotypage à la suite d'évènements mutationnels indépendants (Fenner *et al.,* 2011), observation qui corrobore le fait que le spoligotypage est plus enclin à l'homoplasie que les MIRU-VNTR (Comas *et al.,* 2009). En outre, le spoligotypage présente un faible pouvoir discriminant pour les familles associées à l'absence de gros blocs d'espaceurs, par exemple la lignée « Beijing » (Allix-Béguec *et al.,* 2014). Pour toutes ces raisons, nous recommandons d'effectuer une analyse phylogénétique plus fine des clones de MTBC en circulation qui sont les plus significatifs, avec plusieurs marqueurs génétiques, et de comparer les résultats obtenus aux données existantes au niveau mondial – tâche complexe en soi mais réalisable aujourd'hui grâce à des grandes bases de données internationales offrant cette possibilité.

État des lieux des bases de données sur le génotypage de la TB

Notre connaissance de la TB n'a jamais été aussi grande qu'aujourd'hui, mais qu'en est-il de notre capacité à comparer les données produites à toutes les autres données accumulées au fil des ans ? Sommes-nous réellement capables de comparer instantanément les informations génétiques dont nous disposons sur les souches de MTBC en circulation, avec toutes les données démographiques, cliniques, bactériologiques et épidémiologiques enregistrées à des endroits différents ? La nécessité des bases de données ne fait aucun doute dans un tel contexte, et la conception et la constitution de bases de données, pour le contrôle ou la surveillance de la TB ainsi que d'autres maladies infectieuses, vont certainement constituer une étape déterminante pour atteindre les objectifs du Millénaire pour le développement (OMD) fixés par l'OMS pour 2015 (OMS, 2006). En effet, les bases de données permettent de stocker d'énormes quantités d'informations de manière structurée, facilitant le traitement des données et les recherches et simplifiant le processus décisionnel grâce à la fouille de données guidée par les connaissances. Cependant, à l'ère actuelle du *Big Data*, ces bases de données doivent être constamment mises à jour et entretenues, comme du patrimoine ou des monuments historiques, constituant alors un ensemble de données plus volumineux que les bases de données classiques, si bien que des mesures révolutionnaires doivent être prises en matière de gestion, d'analyse et d'accessibilité des données biologiques (Howe *et al.,* 2008). Depuis deux ans, plusieurs bases de données et outils en ligne ont été mis au point dans le domaine de la TB, plus précisément pour étudier l'évolution et l'épidémiologie moléculaire de la TB ; en voici quelques exemples :

- SpolDB4 et SITVITWEB sont des bases de données de génotypage conçues à l'IPG (Brudey *et al.,* 2006 ; Demay *et al.,* 2012). La dernière version de SITVITWEB est une base de données multimarqueurs qui contient les données de génotypage de 62 582 isolats cliniques correspondant à des patients originaires de 153 pays (105 pays d'isolement). Les techniques de typage utilisées sont les suivantes : (a) spoligotypage, n = 7 105 profils de 58 180 isolats cliniques, regroupés en 2 740 types communs ou SIT (*Spoligotype International Types*) (n = 53 816 isolats cliniques) et 4 364 profils orphelins ; (a) MIRU-VNTR 12 loci, n = 2 379 profils de 8 161 isolats cliniques, regroupés en 847 types communs ou MIT (*MIRU International Types*) (n = 6 626 isolats cliniques) et 1 533 profils orphelins ; (c) séquences exactement répétées en tandem (ETR, *Exact Tandem Repeats*) 5 loci, n = 458 profils de 4 626 isolats cliniques, regroupés en 245 types communs ou VIT (*VNTR International Types*) (n = 4 413 isolats cliniques) et 213 profils orphelins. La base de données SITVITWEB peut être consultée librement sur : http://www.pasteur-guadeloupe.fr:8081/SITVIT_ONLINE
- SpolTools propose des applications en ligne et un outil de visualisation permettant de manipuler et d'analyser les données de sérotypage du MTBC (Reyes *et al.,* 2008 ; Tang *et al.,* 2008). Il comprend également une banque en ligne d'isolats spoligotypés, compilée à partir de la littérature publiée (actuellement 30 ensembles de données contenant 1 179 profils de spoligotypage correspondant à 6 278 isolats). Il permet notamment de dessiner des arbres Spoligoforest qui illustrent les relations d'évolution existant entre différents spoligotypes dans un milieu donné. SpolTools est accessible à l'adresse suivante : http://www.emi.unsw.edu.au/spolTools/
- MIRU-VNTRplus est un outil en ligne conçu pour analyser les données de typage moléculaire liées à la TB, en particulier les formats MIRU-VNTR 12, 15 et 24 loci (Allix-Béguec *et al.,* 2008 ; Weniger *et al.,* 2010). Les outils d'exploration de données permettent de rechercher des souches similaires, de créer des arbres phylogénétiques et des arbres de recouvrement minimaux, et de cartographier des données géographiques. La base de données fournit en outre des résultats détaillés (origine géographique, profils de sensibilité aux médicaments, lignées génétiques et profil de spoligotypage, profils SNP et LSP et empreintes IS6110-RFLP), sur une collection de 186 souches de référence bien caractérisées. MIRU-VNTRplus est accessible à l'adresse suivante : http://www.miru-vntrplus.org/
- TB GIMS (*TB Genotyping Information Management System*) est un système en ligne sécurisé conçu pour améliorer l'accès aux données de génotypage ainsi que leur diffusion à l'échelle nationale aux États-Unis (CDC 2010]. Il stocke et gère les données de génotypage relatives aux patients atteints de TB aux États-Unis ; il permet aux utilisateurs habilités d'envoyer des isolats de MTBC aux laboratoires de génotypage participants et de les suivre en ligne ; il informe immédiatement les laboratoires et les programmes travaillant sur la TB des résultats de génotypage obtenus et des mises à jour effectuées ; il corrèle les données des isolats avec les données de surveillance obtenues au niveau des patients ; il génère des rapports sur les grappes de génotypes, notamment sur la répartition des génotypes à l'échelle nationale ; et il fournit des cartes des grappes de génotypes à l'échelle du pays, des états ou des comtés. Malheureusement, cette base de données n'est pas accessible au public.

Réseaux

- Mbovis.org est une base de données de spoligotypage contenant plus de 1 400 profils appartenant aux lignées de MTBC à délétion RD9 suivantes : *M. africanum, M. bovis* (antilope), *M. microti, M. pinnipedii, M. caprae* et *M. bovis* (Smith et Upton, 2012). Cette base de données peut être consultée sur : http://www.mbovis.org/
- MycoDB.es est une base de données espagnole sur la TB animale (Rodriguez-Campos *et al.,* 2012) qui a été créée pour servir d'outil épidémiologique au niveau national (Espagne). Elle contient 401 profils de spoligotypage différents de 17 273 isolats appartenant à *M. bovis, M. caprae* et *M. tuberculosis*, ainsi qu'une quantité limitée de données de MIRU-VNTR. Malheureusement, l'accès à cette base de données est strictement réservé à l'agence espagnole de santé animale (*Centro de Vigilencia Sanitaria Veterinaria, VISAVET*) : http://www.vigilanciasanitaria.es/mycodb/
- TB-Lineage est un outil en ligne qui permet de classer et d'analyser les génotypes du MTBC en grandes lignées, sur la base du spoligotype et éventuellement du profil MIRU 24 loci (Shabbeer *et al.,* 2012). Il a été conçu et testé avec des données de génotypage obtenues auprès des CDC (*Centers for Disease Control and Prevention, centres épidémiologiques*) d'Atlanta, sur 37 066 isolats cliniques correspondant à 3 198 profils de spoligotypage et 5 430 profils de MIRU-VNTR. Cependant, en l'absence de données de MIRU 24 loci, le système utilise les prédictions effectuées avec un classifieur bayésien naïf sur la base des seules données de spoligotypage. La précision de la classification automatique est supérieure à 99 % avec les données de spoligotypage et de MIRU24 ; avec les spoligotypes seuls, elle est supérieure à 95 %. TB-Lineage est disponible librement sur : http://tbinsight.cs.rpi.edu/run_tb_lineage.html. Ce site propose également un outil qui permet de créer des arbres Spoligoforest afin de visualiser la diversité et la corrélation génétiques des génotypes et des lignées associées.
- Tbvar est une base de données interrogeable qui utilise un pipeline de calcul systématique permettant d'annoter les variantes éventuelles, sur le plan fonctionnel et/ou en matière de résistance aux antibiotiques, par rapport aux données de re-séquençage clinique du MTBC (Joshi *et al.,* 2013). Pour cela, les auteurs ont réanalysé des ensembles de données de re-séquençage correspondant à plus de 450 isolats de MTBC disponibles dans le domaine public afin de produire une carte exhaustive des variomes, constituée de plus de 29 000 variations mononucléotidiques. Cette base de données permet de faire des recherches par lieu des variants (1417019, 3037367, 4222628, etc.) ; gènes (*katG, pncA, gyrA,* etc.) ; RvID (Rv1059, Rv1069c, Rv3693, etc.) ; ou plage de positions dans le génome (10000-15000 ; 30000-35000 ; 80000-85000, etc.) ; elle est accessible à l'adresse suivante : http://genome.igib.res.in/tbvar/
- InTB est une interface en ligne qui permet le stockage et l'analyse intégrés d'informations cliniques et sociodémographiques et de données de typage moléculaire sur la TB (Soares *et al.,* 2013). Ce système permet d'introduire et de télécharger des données de génotypage standards, conjointement avec une large gamme de variables cliniques et sociodémographiques utilisées pour caractériser la maladie. Il permet également de classer les nouveaux isolats dans un ensemble d'isolats bien caractérisés sur la base d'étalons internes, de représenter graphiquement plusieurs types de données et de créer des arbres pour des sous-ensembles de données filtrés combinant données moléculaires et informations cliniques/sociodémographiques. Développés au moyen d'un logiciel *open source*, l'intégralité du code source et des packages prêts à l'emploi sont disponibles à l'adresse suivante : http://www.evocell.org/inTB.

Bases de données de génotypage de la TB élaborées à l'IPG

La première base de données a été créée à l'IPG il y a plus de quinze ans, lorsqu'un stagiaire de deuxième cycle nommé Jérôme Maïsetti a pris l'initiative de saisir dans un tableau Excel les profils de spoligotypage dont nous disposions sur nos propres isolats antillais (n = 218 souches), puis de les mettre en commun avec des données publiées (n = 392 isolats) provenant d'autres pays. Une fois les profils triés, nous nous sommes rendu compte que nous pouvions non seulement identifier des profils prédominants, mais également tracer l'origine des souches et leurs éventuels déplacements. Cette base de données de 610 spoligotypes a été provisoirement nommée « SpolDB1 » et a conduit à la toute première description de 69 grands profils de spoligotypage, permettant de mieux comprendre l'origine et la transmission de la TB (Sola *et al.,* 1999). La création de SpolDB1 a été suivie du lancement de SpolDB2, contenant des données sur 3 319 isolats (Sola *et al.,* 2001), puis de SpolDB3, sur 13 008 isolats regroupés en 813 types communs (contenant 11 708 isolats) et 1 300 profils orphelins (Filliol *et al.,* 2002 ; 2003). Plus récemment, l'élaboration de la quatrième version SpolDB4 au format MySQL en 2006, n = 39 295 isolats cliniques (Brudey *et al.,* 2006), et de SITVITWEB en 2012, n = 62 582 isolats cliniques (Demay *et al.,* 2012), a dressé un portrait plus précis de la phylogéographie des lignées génotypiques de MTBC en circulation dans le monde. À ce jour, la dernière version de notre base de données, SITVIT2 (n = 111 635 isolats), contient les données de génotypage d'environ deux fois plus de souches que la version précédente ; elle sera publiée en 2014. Il convient de souligner que ces deux versions récentes sont des bases de données multimarqueurs au format MySQL qui contiennent des données de spoligotypage et de typage MIRU-VNTR, limitées aux MIRU 12 loci dans SITVITWEB ; et des données de MIRU-VNTR 12, 15 et 24 loci dans SITVIT2. Sur ces dernières versions, une interface en ligne permet à l'utilisateur de rechercher des souches dans la base de données au moyen de critères tels que l'année, le pays d'isolement, le pays d'origine ou le nom du chercheur ; cette interface permet en outre d'effectuer des recherches mixtes dans SITVIT2, permettant d'obtenir les données de génotypage de certaines souches, conjointement avec leur répartition géographique, ainsi que les données disponibles sur la résistance aux antibiotiques ou les caractéristiques démographiques et épidémiologiques.
Grâce à l'élaboration de ces bases de données successives, l'IPG a considérablement amélioré sa connaissance du génotypage et de la phylogénie/phylogéographie de la TB dans le monde. Les différentes versions de ces bases de données ont permis la réalisation d'un nombre significatif d'études bilatérales et multilatérales, comme le prouve le très grand nombre de citations des bases de données successives (interrogation effectuée sur Google Scholar le 31 mars 2014) : SpolDB3 (Filliol *et al.,* 2002 ; 2003), 180 + 220, soit un total de 400 citations ; SpolDB4 (Brudey *et al.,* 2006), 661 citations ; et SITVITWEB (Demay *et al.,* 2012), 67 citations, bien que cette base de données soit en ligne depuis seulement un an. Nous sommes donc certains que la prochaine version

Réseaux

SITVIT2 trouvera elle aussi sa place auprès de la communauté scientifique en tant qu'outil utile non seulement pour la génétique moléculaire des populations, la démographie historique et la modélisation épidémiologique de la TB, mais aussi pour les analyses génétiques fondamentales.

Présentation de SITVIT2

Les fonctionnalités du futur site Internet SITVIT2 seront améliorées par rapport à la version actuelle de SITVITWEB, tant sur le plan numérique qu'au niveau de l'interface de recherche ; la désignation des lignées demeurera quant à elle pratiquement inchangée, à quelques exceptions près. À l'heure où nous écrivons ces lignes, SITVIT2 contient au total 111 635 isolats cliniques de MTBC prélevés chez des patients originaires de 169 pays. Pour la collecte des données, nous avons enrichi la base de données avec les résultats de génotypage obtenus à l'IPG ou transmis par différents laboratoires co-investigateurs et partenaires, ou avec des données extraites d'études publiées (Demay *et al.,* 2012). Le site Internet a été développé en langage JSP (*Java Server Pages*) et est hébergé sur un serveur d'applications Apache Tomcat gratuit (http://tomcat.apache.org), dans les locaux de l'IPG. La technologie Java a été utilisée comme décrit précédemment (Demay *et al.,* 2012). Comme sur les versions précédentes, la description des caractères génétiques des isolats cliniques consultable dans SITVIT2 repose sur une clé d'identification unique (IsoNumber) qui reprend les informations relatives au pays d'isolement, le code du laboratoire, l'année d'isolement, un code pour la résistance aux médicaments (0 à 4) et un numéro d'isolat unique attribué par le laboratoire ou l'hôpital participant. Conformément aux principes éthiques relatifs au traitement électronique des données, cette procédure permet d'attribuer un numéro anonyme que seule la personne ayant fourni les données (le laboratoire de microbiologie qui a transmis les données), et aucun autre utilisateur, peut décoder pour remonter jusqu'aux informations relatives au patient. SITVIT2 utilise le système de marquage automatique de SpolDB4 et de SITVITWEB, qui attribue un numéro SIT à chaque spoligotype présent dans 2 souches ou plus de la base de données, et un numéro MIT à chaque profil MIRU présent dans 2 souches ou plus. Les numéros MIT relatifs aux formats MIRU-VNTR 12, 15 et 24 loci sont respectivement dénommés 12-MIT, 15-MIT et 24-MIT, tandis que ceux réservés aux ETR sont dénommés VIT. Le terme « orphelin » fait quant à lui référence aux profils décrits pour un seul isolat qui ne correspond à aucun des profils enregistrés dans la banque de la base de données SITVIT2.
Parmi les fonctionnalités du site Internet, on peut notamment citer plusieurs outils de recherche pour la conversion des formats de spoligotype (de binaire en octal et vice versa), l'envoi et l'analyse des données de spoligotypage et de MIRU sous différents formats, ainsi que la recherche par critères pour les marqueurs, seuls ou combinés, l'année et le pays d'isolement/d'origine, le nom du chercheur, des cartes de répartition géographique et les informations connexes sur la démographie, l'épidémiologie et la résistance aux antituberculeux. Il est donc possible d'effectuer des recherches individuelles ou groupées en téléchargeant un fichier Excel correctement formaté (modèle disponible sur le site). Pour toutes les recherches, l'utilisateur pourrait s'attendre à obtenir un rapport détaillé sur les marqueurs (numéros SIT et MIT), les lignées phylogénétiques et les informations connexes disponibles dans la base de données sous un format anonymisé. Parmi les améliorations prévues, il sera possible de visualiser les correspondances de nomenclature des profils MIT dans SITVIT2, d'après la nomenclature de MIRU-VNTRplus (on notera qu'une comparaison des lignées entre les deux bases de données n'a pas révélé de différences significatives ; résultats non présentés).

Grandes lignées phylogéniques de SITVIT2

Dans SITVIT2, les souches sont classées en grands clades phylogénétiques, selon les signatures décrites précédemment, comprenant plusieurs membres du MTBC (AFRI, *M. africanum* ; BOV, *M. bovis* ; CANETTII, *M. canettii* ; MICROTI, *M. microti* ; PINI, *M. pinnipedii*), ainsi que *M. tuberculosis* stricto sensu, à savoir les clades : « Beijing », CAS (Central Asian), EAI (East-African Indian), « Haarlem/Ural », LAM (Latin-American-Mediterranean), « Cameroon », « Turkey », « Manu », X (lignée à faible de nombre de bandes IS6110) et le clade T, mal défini. Il convient de souligner que certains spoligotypes préalablement classés dans les sous-lignées H3/H4 au sein de la famille « Haarlem » ont récemment été reclassés dans le clade « Ural » (Mokrousov, 2012) ; il s'agit notamment de profils appartenant à la sous-lignée H4 renommés « Ural 2 » et de certains profils préalablement classés dans la sous-lignée H3, mais présentant une autre signature spécifique (présence de l'espaceur 2, absence des espaceurs 29 à 31 et 33 à 36), aujourd'hui renommés « Ural 1 ». De surcroît, deux sous-lignées LAM ont récemment été élevées au niveau de lignée indépendante : LAM10-CAM pour la lignée « Cameroon » (Koro Koro *et al.,* 2013) et LAM7-TUR pour la lignée « Turkey » (Abadia *et al.,* 2010 ; Kisa *et al.,* 2012). Nous avons conservé cette nomenclature pour les lignées génotypiques identifiées par spoligotypage, dans la mesure où elle s'est déjà révélée utile pour des études d'épidémiologie moléculaire locales ou internationales, ainsi que pour suivre l'évolution et la génétique quantitative des bacilles tuberculeux à l'échelle mondiale. On notera que la répartition des isolats cliniques de SITVIT2 est étudiée tant à l'échelle nationale qu'au niveau macrogéographique, par sous-régions, conformément à la classification des Nations Unies (voir http://unstats.un.org/unsd/methods/m49/m49regin.htm) ; régions : AFRI (Afrique), AMER (Amériques), ASIA (Asie), EURO (Europe) et OCE (Océanie), subdivisées en : E (est), M (milieu), C (centre), N (nord), S (sud), SE (sud-est) et W (ouest). Selon ce système de classification, les Antilles (CARIB) appartiennent aux Amériques, tandis que l'Océanie est subdivisée en 4 sous-régions : AUST (Australasie), MEL (Mélanésie), MIC (Micronésie) et POLY (Polynésie). Il convient de signaler que la Russie s'est vue attribuer une nouvelle sous-région (Asie du Nord), au lieu d'être intégrée à l'Europe de l'Est.
Les lecteurs sont invités à se reporter au **tableau 2** pour une comparaison rapide de SITVITWEB et SITVIT2 et des lignées phylogénétiques correspondantes dans les 2 versions, ainsi qu'à la **figure 1** qui illustre l'évolution des souches entre les 2 versions, par sous-région géographique. L'augmentation la plus significative entre les 2 versions concerne l'Europe du Sud et l'Asie de l'Est, suivies de l'Amérique Centrale et du Sud, de l'Afrique du Nord, de l'Est et de l'Ouest, et de l'Europe du Nord et de l'Ouest **(figure 1)**. Quant aux lignées, la proportion du

Réseaux

génotype « Beijing » n'a pas varié de manière significative entre SITVITWEB et SITVIT2 (représentant respectivement 9,84 % contre 9,72 % des isolats au niveau mondial ; tableau 2) ; ce génotype est prédominant en Asie et sa proportion est significative en Amérique du Nord, en Afrique du Sud et en Australasie. Les proportions sont similaires pour les lignées CAS (3,69 % contre 3,91 %), « Cameroon » (anciennement LAM10-CAM : 1,04 % contre 0,98 %), « Turkey » (anciennement LAM7-TUR : 0,59 % contre 0,53 %) et « Manu » (1,08 % contre 0,95 %). Il convient cependant de noter que la plus forte augmentation concerne les souches de *M. bovis* (10,36 % contre 23,06 %), ce qui souligne le potentiel de SITVIT2 pour l'étude de l'épidémiologie de *M. tuberculosis* ainsi que l'étude de la TB bovine.

Tableau 2. Comparaison rapide des bases de données SITVITWEB et SITVIT2 et des grandes lignées phylogénétiques correspondantes du MTBC.

Major Lineages*	SITVITWEB * (n=62582)		SITVIT2 (n=111635)	
	Nb	%	Nb	%
Beijing	6159	9.84	10850	9.72
AFRI	695	1.11	965	0.86
BOV	6486	10.36	25741	23.06
CANETTII	12	0.02	12	0.01
CAS	2480	3.96	4362	3.91
EAI	4674	7.47	6617	5.93
Haarlem/Ural	7058	11.28	10580	9.48
LAM	8042	12.85	12245	10.97
Cameroon (previously LAM10-CAM)	650	1.04	1095	0.98
Turkey (previously LAM7-TUR)	370	0.59	593	0.53
Manu	675	1.08	1064	0.95
MICROTI	29	0.05	29	0.03
PINI	152	0.24	159	0.14
S	1151	1.84	1606	1.44
T	12038	19.24	17947	16.08
X	4088	6.53	4683	4.19

Cartes de répartition mondiale de SITVIT2

La carte de répartition mondiale des grandes lignées de SITVIT2, représentée sur la **figure 2**, illustre les géospécificités phylogéographiques mondiales des isolats de MTBC, donnant une vue d'ensemble de la situation en 2014. Bien que ces géospécificités aient déjà été évoquées dans des études précédentes fondées à la fois sur le spoligotypage (Brudey *et al.,* 2006 ; Demay *et al.,* 2012) et sur le typage LSP (Gagneux *et al.,* 2006), cette carte confirme le profil de répartition et les spécificités actuels grâce aux données conservées dans la base de données SITVIT2. De plus, la reclassification de certaines lignées qui n'apparaissaient pas dans les versions précédentes, comme « Ural », montre que cette lignée est très présente en Russie, en Asie centrale, en Asie du Sud et en Asie de l'Ouest, ainsi qu'en Finlande (Europe du Nord), qui partage une frontière commune et des liens privilégiés avec la Russie, en ayant notamment joué le rôle de zone tampon dans une succession de guerres entre la Russie et la Suède au cours du 18e siècle (http://www.historyworld.net/wrldhis/PlainTextHistories.asp?historyid=ad02). Nous avons par ailleurs confirmé le réétiquetage de LAM7-TUR comme lignée « Turkey » et celui de LAM10-CAM en lignée « Cameroon », grâce à leur répartition phylogéographique dans SITVIT2 **(figure 2)**. Il est important de souligner que la lignée « Turkey » progresse dans les pays d'Europe de l'Est (représentant près de 6 % de toutes les souches de MTBC d'Europe de l'Est en 2014, contre moins de 2 % jusqu'en 2006 ; résultats non présentés). Il convient également de noter la diminution continue de la proportion de la lignée AFRI en Afrique de l'Ouest, qui représentait environ 37 % de toutes les souches de MTBC dans cette sous-région jusqu'en 2006, contre 29 % en 2014. Cette observation corrobore notre précédente hypothèse selon laquelle les souches de *M. africanum* ancestrales d'Afrique de l'Ouest sont peu à peu remplacées par des lignées récentes de MTBC comme « Cameroon » ou d'autres lignées euro-américaines (Groenheit *et al.,* 2011). Enfin et surtout, bien que les proportions globales de CAS n'aient pas beaucoup évolué entre SITVITWEB et SITVIT2 (3,69 % contre 3,91 %), on observe une augmentation en Asie de l'Ouest (12 % contre 18 %) et en Afrique de l'Est (10 % contre 15 %).

Recherche de corrélations fortes entre lignées phylogéniques et caractéristiques démographiques et épidémiologiques dans SITVIT2

Un utilisateur pourra obtenir des données phylogénétiques associées à un certain nombre de paramètres différents tels que l'incidence de la maladie représentée sur des cartes disponibles auprès de l'OMS, la démographie (âge, sex-ratio) ou d'autres caractéristiques dans différentes sous-régions, afin de mettre en évidence les spécificités de chaque pays, région ou population. Si l'on considère que la propagation des clones MDR et XDR de TB en population générale représente une menace de premier plan en matière de lutte antituberculeuse, l'intérêt d'une telle base de données est également de donner une vue d'ensemble de la situation actuelle et de localiser toute émergence éventuelle pour les autorités de santé publique. Ce géoréférencement de nos données, au moyen d'une API Google, peut déjà former un SIG constituant un pendant bactériologique efficace de l'Atlas mondial des maladies infectieuses de l'OMS (http://apps.who.int/globalatlas/), réunissant en une seule plate-forme électronique l'analyse et l'interprétation des données de génotypage et des informations relatives à la démographie, aux conditions socioéconomiques ou aux facteurs environnementaux. Nous pensons qu'une telle cartographie devrait idéalement être associée à des outils et logiciels statistiques et bio-informatiques appropriés afin de mieux décrire le paysage génétique de la TB. Outre les outils

*Les souches sont classées en grands clades phylogénétiques, selon les signatures décrites précédemment (Demay et al., 2012) ; elles comprennent plusieurs membres du MTBC (AFRI, *M. africanum* ; BOV, *M. bovis* ; CANETTII, *M. canettii* ; MICROTI, *M. microti* ; PINI, *M. pinnipedii*), ainsi que des lignées/sous-lignées de M. tuberculosis stricto sensu (les sous-lignées ne sont pas indiquées) : le clade « Beijing », le clade CAS, le clade EAI, les clades « Haarlem/Ural », le clade LAM, les lignées « Cameroon » et « Turkey », la famille « Manu », le clade X à faible nombre de bandes IS6110 et le clade T, mal défini.

Réseaux

Figure 1. (A) Répartition mondiale des isolats de MTBC enregistrés dans SITVITWEB (cercles bleus) et SITVIT2 (cercles verts), le nombre d'isolats par sous-région étant indiqué à l'intérieur de chaque cercle. (B) Évolution progressive d'une série de bases de données (de SpolDB2 à SITVIT2), selon le nombre d'isolats de MTBC contenus dans chaque base de données.

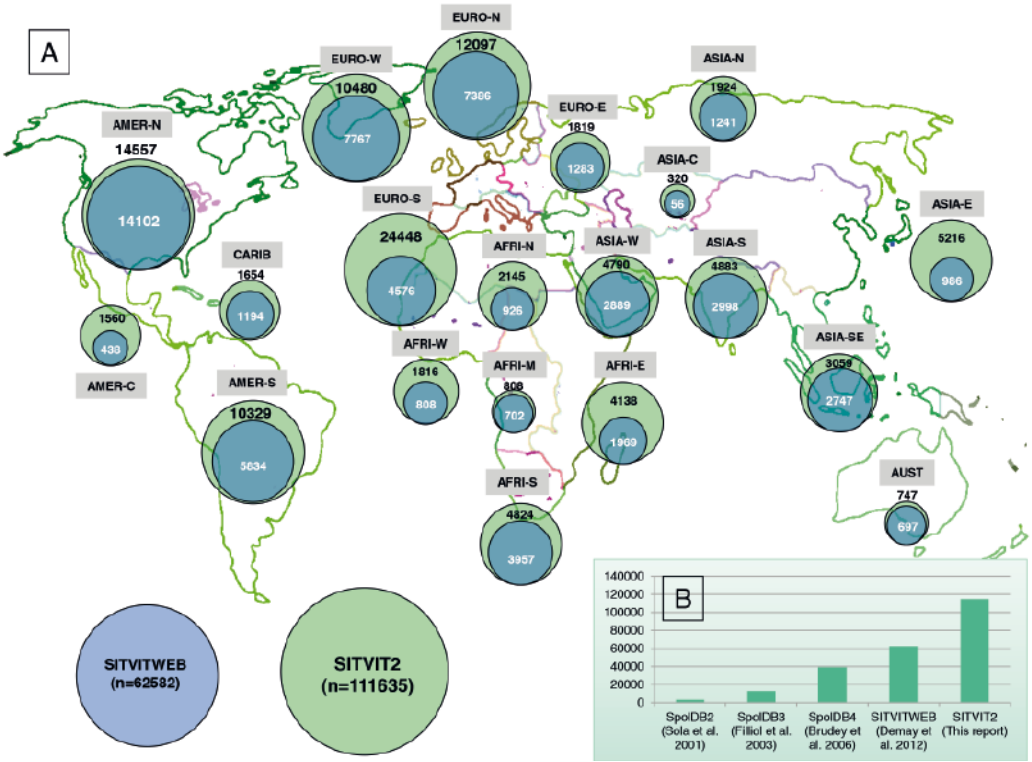
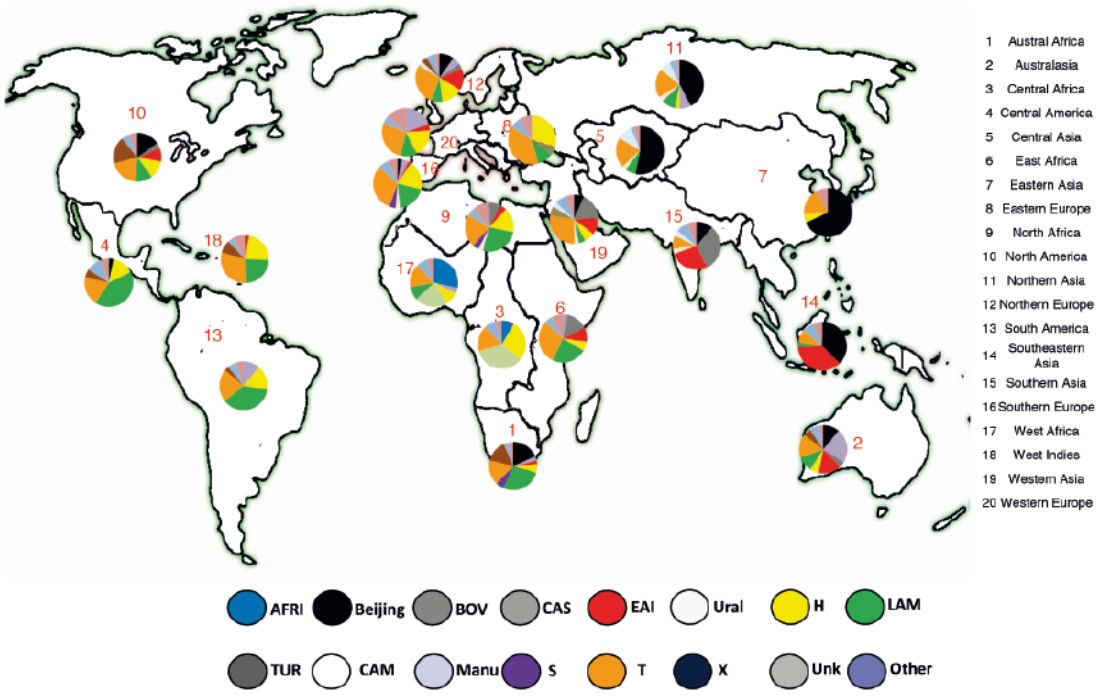


Figure 2 : Répartition mondiale des lignées contenues dans la base de données SITVIT2.



Réseaux

et logiciels décrits dans les études précédentes (Brudey *et al.,* 2006 ; Demay *et al.,* 2012), nous utilisons aussi les outils suivants pour l'analyse courante des données de SITVIT2 :
(a) Le logiciel STATA, version 12, pour des analyses descriptives et univariées.
(b) Le logiciel R, version 2.14.1, pour calculer l'Odds Ratios (OR) et les bornes de l'intervalle de confiance (IC) à 95 %.
(c) Le test du chi carré de Pearson et le test exact de Fisher pour comparer les fortes corrélations existant entre les données de génotypage (types communs/lignées) et les caractéristiques démographiques, épidémiologiques ou socioéconomiques (des valeurs p inférieures à 0,05 étant considérée statistiquement significatives).
(d) Des arbres de recouvrement minimal (ARM) sont construits à partir des données de génotypage (spoligotypes, profils MIRU) à l'aide du logiciel BioNumerics, version 6.6 (Applied Maths, Sint-Martens-Latem, Belgique). Il s'agit de graphes non orientés associés sur lesquels tous les profils sont reliés, avec le moins de liens possible entre les voisins les plus proches.
(e) Le logiciel SpolTools (http://www.emi.unsw.edu.au/spolTools) permet de dessiner des arbres Spoligoforest sur la base de l'algorithme de Fruchterman et Reingold ou d'une disposition hiérarchique (Reyes *et al.,* 2008 ; Tang *et al.,* 2008). On notera que, contrairement aux ARM, les arbres Spoligoforest sont des graphes orientés (et pas forcément associés) qui représentent les relations d'évolution existant entre les profils de spoligotypage des ascendants et des descendants.
(f) Le logiciel GraphViz, disponible sur http://www.graphviz.org (Ellson *et al.,* 2002), pour colorier les arbres Spoligoforest en fonction des lignées.
(g L'application WebLogo, version 2.8.2, disponible sur http://weblogo.berkeley.edu/ (Schneider and Stephens, 1990 ; Crooks *et al.,* 2004), pour évaluer et visualiser la diversité allélique des profils de spoligotypage en fonction de leurs lignées associées. Cette méthode de représentation adaptée au spoligotypage sur 43 espaceurs a été dénommée « Spoligologos » (Driscoll *et al.,* 2002). WebLogos comprend des piles de symboles qui permettent de représenter graphiquement la diversité comparative observée pour chaque espaceur de spoligotypage – une pile pour chacun des 43 espaceurs. La hauteur totale d'une pile indique le degré de conservation d'un espaceur donné (la lettre « n » indique la présence d'un espaceur et la lettre « o » son absence). Si un espaceur est systématiquement présent ou absent en une position donnée parmi les 43 disponibles (en d'autres termes, si 100 % des souches conservent la même présence ou la même absence en une position donnée), cela correspond à 4 bits ; la hauteur de chaque symbole dans la pile indique quant à elle la fréquence relative des espaceurs absents/présents en cette position.

Exemples d'études récentes utilisant la base de données SITVIT2

Il est bien entendu impossible de citer toutes les études réalisées avec SITVIT2 ; on peut cependant mentionner quelques exemples de travaux récents qui portent sur les corrélations entre lignées phylogénétiques et paramètres démographiques et épidémiologiques :
(a) Carte de répartition géographique des lignées de MTBC identifiées par spoligotypage dans plusieurs sous-régions d'Afrique et forte spécificité phylogéographique de *M. africanum* pour l'Afrique de l'Ouest, la Guinée-Bissau étant l'épicentre (Groenheit *et al.,* 2011).
(b) Preuve que les souches de MTBC potentiellement responsables de l'épidémie de TB qui a touché la Suède il y a un siècle appartenaient à un groupe de souches PGG2/3 d'évolution récente étroitement apparentées, limité à la Suède et aux pays immédiatement voisins (Groenheit *et al.,* 2012).
(c) Analyse des corrélations phylogénétiques entre le MTBC et la résistance aux antituberculeux au Pérou, suggérant une épidémie nosocomiale clonale prolongée des isolats MDR chez des patients infectés par le VIH (Sheen *et al.,* 2013).
(d) Cartographie phylogéographique de la TB en Finlande, révélant une forte ressemblance entre la structure de la population mondiale du MTBC et une structure décrite en Suède, en particulier la prédominance de la famille euro-américaine chez les patients âgés atteints de TB ; la principale différence résidant dans la lignée « Ural », qui a été observée en grandes proportions dans les cas d'origine finlandaise (et qui a aussi été observée en Russie, en Lettonie et en Estonie), mais pas en Suède (Smit *et al.,* 2013).
(e) Au niveau mondial, l'utilisation de la base de données SITVIT2 nous a permis de découvrir une corrélation plus significative entre les lignées X et LAM et une sérologie positive au VIH ; valeur $p < 0,0001$ (article sous presse).
(f) Plusieurs études ont mis en évidence la corrélation existant entre la lignée « Beijing » et une résistance excessive aux médicaments, y compris la TB MDR/XDR (van Soolingen *et al.,* 1995 ; Glynn *et al.,* 2002 ; Parwati *et al.,* 2010). Ainsi, nous avons récemment analysé les corrélations phylogénétiques (MTBC divisé en 2 groupes : « Beijing » et autres lignées) avec la résistance aux antibiotiques (quantifiée comme – pansusceptible, MDR, XDR ou autre) en utilisant la base de données SITVIT2 (Couvin et Rastogi, 2014). La **figure 3A** illustre la répartition de la résistance dans différentes sous-régions. Bien que la proportion de souches résistantes soit beaucoup plus élevée pour les souches « Beijing » que pour les autres souches à l'échelle mondiale, nous observons des variations importantes quant à la répartition mondiale de la résistance aux antituberculeux. La résistance est fortement corrélée aux souches « Beijing » par rapport aux autres souches en Russie, en Asie du Sud, en Asie du Sud-Est et dans les pays européens, mais pas en Amérique, en Asie de l'Ouest, en Chine et au Japon. Si l'on analyse l'évolution de la résistance dans le temps pour les souches « Beijing » (de 1998 à 2011, **figure 3B**), on observe une progression continue de la proportion de souches MDR et XDR (et une diminution relative des souches pansusceptibles) à l'échelle mondiale depuis 2003. Enfin et surtout, nous observons également qu'un profil de spoligotypage rare mais émergent, SIT190/Beijing, présente une corrélation plus significative avec la TB MDR que le profil SIT1/Beijing classique (valeur $p < 0,0001$).
(g) Concernant les informations obtenues grâce à un arbre Spoligoforest, nous souhaitons citer une étude récente menée à Bagdad en Iraq (Mustafa Ali *et al.,* 2014). Les résultats obtenus, sur un total de 270 isolats de MTBC, ont montré que 2 profils spécifiques, SIT1144/T1 et SIT309/CAS1-Delhi, prédominaient dans cette étude (6,3 % pour chaque profil). Les relations d'évolution existant entre les isolats iraquiens, visibles sur un arbre Spoligoforest avec disposition hiérarchique (**figure 4**), montrent clairement que la majeure partie de la TB observée dans l'Iraq d'après-guerre se limite à 2 groupes de souches de MTBC phylogénétiquement apparentées, appartenant aux lignées génotypiques T et CAS.
(h)Concernant les informations obtenues grâce au

Réseaux

Figure 3. Caractéristiques de résistance aux médicaments des lignées de *M. tuberculosis* « Beijing » et « non-Beijing » (A) et évolution de la résistance aux médicaments dans les isolats « Beijing » entre 1998 et 2011 (B)

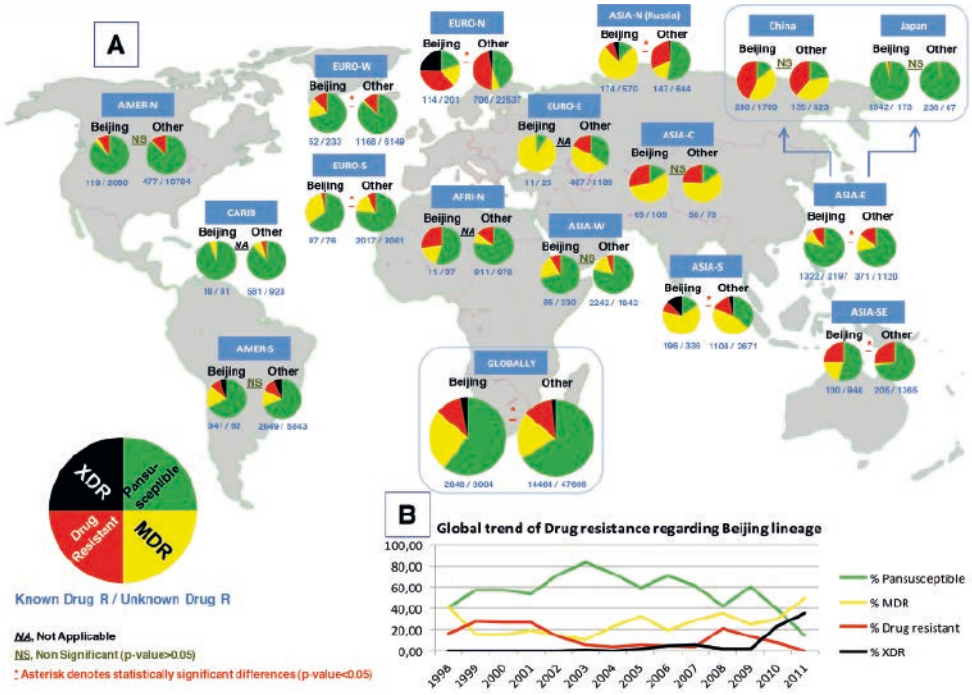
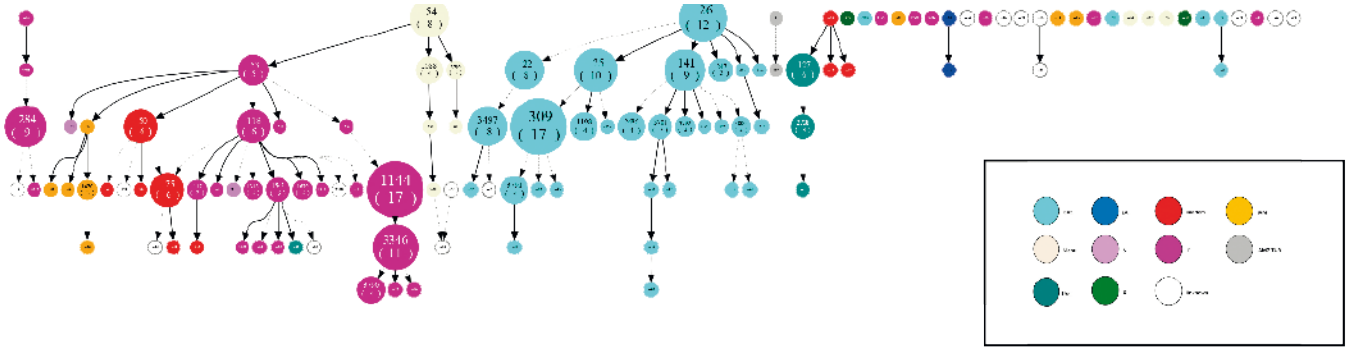
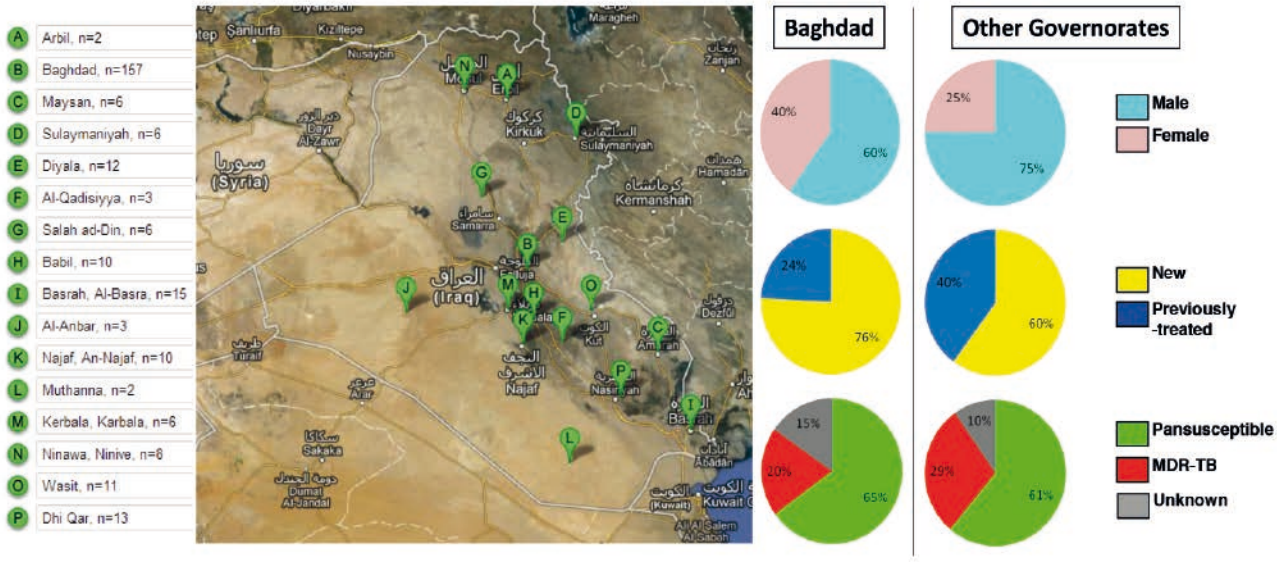


Figure 4. Graphique Spoligoforest, des parents aux descendants, lors d'une étude menée à Bagdad en Iraq (n = 270 isolats), avec une disposition hiérarchique. Sur cet arbre, chaque profil de spoligotypage est symbolisé par un nœud dont la taille est proportionnelle au nombre total d'isolats ayant ce profil. Les modifications (perte d'espaceurs) sont symbolisées par des flèches, entre les nœuds, qui pointent sur les spoligotypes des descendants. L'heuristique utilisée choisit une seule flèche entrante de poids maximal, au moyen d'un modèle de Zipf. Les traits pleins noirs relient les profils très similaires, à savoir ceux qui n'ont perdu qu'un seul espaceur (le poids maximal étant 1,0), tandis que les traits discontinus relient les poids compris entre 0,5 et 1, et les pointillés les poids inférieurs à 0,5. Il convient de noter que SIT309/CAS1-Delhi et SIT1144/T1 sont les deux plus gros nœuds (n = 17), suivis de SIT26/CAS1-Delhi (n = 12), SIT3346/T1 (n = 11) et SIT25/CAS1-Delhi (n = 10), qui sont les autres profils prédominants dans notre étude. Enfin, les isolats orphelins (doubles cercles) apparaissent principalement en positions terminales sur l'arbre ou sont des souches isolées sans liens avec les autres souches (figure élaborée à partir des données de Mustafa Ali et al., 2014).



Réseaux

Figure 5. Capture d'écran de Google Maps montrant la répartition des souches de MTBC en fonction de la ville d'isolement, à Bagdad et dans d'autres villes d'Iraq (n = 270 isolats), et comparaison des données en fonction de l'origine des patients et du rapport de masculinité, de la phase de traitement ou de la résistance aux médicaments à Bagdad et dans les autres gouvernorats (figure élaborée à partir des données de Mustafa Ali et al., 2014).



géoréférencement des données de génotypage avec une API Google, nous pouvons citer la même étude (Mustafa Ali *et al.,* 2014). Comme le montre la **figure 5**, il existe des différences significatives entre la ville de Bagdad et les autres villes iraquiennes, en matière d'informations démographiques et de données de résistance aux médicaments. En effet, avec une sex-ratio hommes/femmes de 1,49 à Bagdad, contre 3,04 dans les autres gouvernorats, la proportion de patients féminins est nettement supérieure dans la ville de Bagdad (valeur p = 0,009 ; Odds Ratio = 0,49 et IC 95 % [0,28 ; 0,86]). Le rapport entre les cas traités pour la première fois et les cas retraités varie considérablement d'un groupe à l'autre, la proportion de patients retraités étant plus élevée dans les autres gouvernorats iraquiens (valeur p = 0,007 ; OR = 2,1, IC 95 % [1,19 ; 3,62]). Enfin, le taux de TB MDR est supérieur dans les gouvernorats autres que Bagdad (p > 0,12 ; différence non statistiquement significative).

Conclusions

La collection de bases de données élaborées à l'IPG permet de disposer d'une vue d'ensemble de la situation mondiale en matière de TB. Avec la base de données SITVIT2 récemment mise à jour, nous avons par ailleurs pu décrire la circulation actuelle des souches de MTBC, sur la base de marqueurs de génotypage élargis, et nous avons également mis en évidence de fortes corrélations entre les lignées phylogénétiques de MTBC et les caractéristiques démographiques et épidémiologiques. Dans l'idéal, les prochaines améliorations devraient comprendre l'intégration d'autres marqueurs, comme les RD/LSP et les SNP, ainsi que des futures informations produites grâce au séquençage de nouvelle génération. Bien que de nombreuses caractéristiques évolutives et pathobiologiques de l'épidémie de TB actuelle restent à découvrir, la nouvelle collection de bases de données SITVIT constitue un outil de référence pour améliorer la surveillance épidémiologique de la TB et la lutte antituberculeuse.

Remerciements

Nous remercions les plus de cinq cents chercheurs qui ont fourni les données qui ont permis de constituer les bases de données SpolDB4, SITVITWEB et SITVIT2 (dont 190 ont fourni les données de génotypage de plus de 100 souches). Il convient de souligner que chaque souche et les données de génotypage respectives qui figurent dans chacune de ces bases de données font directement référence au chercheur concerné (liste exhaustive disponible sur demande). Nous remercions particulièrement Thierry Zozio, Julie Millet, Véronique Hill et Élisabeth Streit pour leurs réflexions très utiles. DC a reçu une bourse de doctorat du Fonds social européen, par l'intermédiaire du Conseil régional de Guadeloupe.

Bibliographie

1. Abadia E, Zhang J, dos Vultos T, Ritacco V, Kremer K, Aktas E, Matsumoto T, Refregier G, van Soolingen D, Gicquel B, Sola C. (2010). Resolving lineage assignation on *Mycobacterium tuberculosis* clinical isolates classified by spoligotyping with a new high-throughput 3R SNPs based method. *Infect Genet Evol*. 2010; 107: 1066–1074.

2. Allix-Béguec C, Harmsen D, Weniger T, Supply P, Niemann S. 2008.

Réseaux

Evaluation and strategy for use of MIRU-VNTRplus, a multifunctional database for online analysis of genotyping data and phylogenetic identification of *Mycobacterium tuberculosis* complex isolates. *J Clin Microbiol*. 46: 2692-2699.

3. Allix-Béguec C, Wahl C, Hanekom M, Nikolayevskyy V, Drobniewski F, Maeda S, Campos-Herrero I, Mokrousov I, Niemann S, Kontsevaya I, Rastogi N, Samper S, Sng LH, Warren RM, Supply P. 2014. Proposal of a consensus set of hypervariable mycobacterial interspersed repetitive-unit-variable-number tandem-repeat loci for subtyping of *Mycobacterium tuberculosis* Beijing isolates. *J Clin Microbiol*. 52: 164-172.

4. Barry CE 3rd, Boshoff HI, Dartois V, Dick T, Ehrt S, Flynn J, Schnappinger D, Wilkinson RJ, Young D. 2009. The spectrum of latent tuberculosis: rethinking the biology and intervention strategies. *Nat Rev Microbiol*. 7: 845-855.

5. de Boer AS, Borgdorff MW, de Haas PE, Nagelkerke NJ, van Embden JD, van Soolingen D. 1999. Analysis of rate of change of IS6110 RFLP patterns of *Mycobacterium tuberculosis* based on serial patient isolates. *J Infect Dis*. 180: 1238-1244.

6. Brosch R, Gordon SV, Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, Garnier T, Gutierrez C, Hewinson G, Kremer K, Parsons LM, Pym AS, Samper S, van Soolingen D, Cole ST. 2002. A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proc Natl Acad Sci USA*. 99: 3684-3689.

7. Brudey K, Driscoll JR, Rigouts L, Prodinger WM, Gori A, Al-Hajoj SA, Allix C, Aristimuño L, Arora J, Baumanis V, Binder L, Cafrune P, Cataldi A, Cheong S, Diel R, Ellermeier C, Evans JT, Fauville-Dufaux M, Ferdinand S, Garcia de Viedma D, Garzelli C, Gazzola L, Gomes HM, Guttierez MC, Hawkey PM, van Helden PD, Kadival GV, Kreiswirth BN, Kremer K, Kubin M, Kulkarni SP, Liens B, Lillebaek T, Ho ML, Martin C, Martin C, Mokrousov I, Narvskaïa O, Ngeow YF, Naumann L, Niemann S, Parwati I, Rahim Z, Rasolofo-Razanamparany V, Rasolonavalona T, Rossetti ML, Rüsch-Gerdes S, Sajduda A, Samper S, Shemyakin IG, Singh UB, Somoskovi A, Skuce RA, van Soolingen D, Streicher EM, Suffys PN, Tortoli E, Tracevska T, Vincent V, Victor TC, Warren RM, Yap SF, Zaman K, Portaels F, Rastogi N, Sola C. 2006. *Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. BMC Microbiol. 6: 23.

8. CDC. 2010. Launch of TB Genotyping Information Management System (TB GIMS). *Morbidity and Mortality Weekly Report* (MMWR) March 19, 59(10); 300.

9. Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, Parkhill J, Malla B, Berg S, Thwaites G, Yeboah-Manu D, Bothamley G, Mei J, Wei L, Bentley S, Harris SR, Niemann S, Diel R, Aseffa A, Gao Q, Young D, Gagneux S. 2013. Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. *Nat Genet*. 45: 1176-1182.

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

11. Couvin D, Rastogi N. (2014). Tuberculosis – a global emergency: tools and methods to monitor, understand, and control the epidemic with specific example of the Beijing lineage. Tuberculosis (Edinb). Submitted for publication.

12. Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res*. 14: 1188-1190.

13. Dale JW, Brittain D, Cataldi AA, Cousins D, Crawford JT, Driscoll J, Heersma H, Lillebaek T, Quitugua T, Rastogi N, Skuce RA, Sola C, Van Soolingen D, Vincent V. 2001. Spacer oligonucleotide typing of bacteria of the *Mycobacterium tuberculosis* complex: recommendations for standardised nomenclature. *Int J Tuberc Lung* Dis. 5: 216-219.

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

15. Driscoll JR, Bifani PJ, Mathema B, McGarry MA, Zickas GM, Kreiswirth BN, Taber HW. 2002. Spoligologos: a bioinformatic approach to displaying and analyzing *Mycobacterium tuberculosis* data. *Emerg Infect Dis*. 8(11): 1306-9.

16. Ellson J, Gansner E, Koutsofios L, North SC, Woodhull G. 2002. Graphviz – Open Source Graph Drawing Tools. In: Mutzel P, Jünger M, Leipert S (Editors), Heidelberg: Springer-Verlag Berlin. pp. 483-484.

17. van Embden JDA, Cave MD, Crawford JT, Dale JW, Eisenach KD, Gicquel B, Hermans P, Martin C, McAdam R, Shinnick TM, Small PM. 1993. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. *J Clin Microbiol.* 31, 406-409.

18. Fang Z, Kenna DT, Doig C, Smittipat DN, Palittapongarnpim P, Watt B, Forbes KJ. 2001. Molecular evidence for independent occurrence of IS6110 insertions at the same sites of the genome of *Mycobacterium tuberculosis* in different clinical isolates. *J Bacteriol.* 183: 5279-5284.

19. Fenner L, Malla B, Ninet B, Dubuis O, Stucki D, Borrell S, Huna T, Bodmer T, Egger M, Gagneux S. (2011) "Pseudo-Beijing": evidence for convergent evolution in the direct repeat region of *Mycobacterium tuberculosis*. *PLoS One*. 6: e24737.

20. Filliol I, Motiwala AS, Cavatore M, Qi W, Hazbón MH, Bobadilla del Valle M, Fyfe J, García-García L, Rastogi N, Sola C, Zozio T, Guerrero MI, León CI, Crabtree J, Angiuoli S, Eisenach KD, Durmaz R, Joloba ML, Rendón A, Sifuentes-Osornio J, Ponce de León A, Cave MD, Fleischmann R, Whittam TS, Alland D. 2006. Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for *a minima*l standard SNP set. *J Bacteriol*. 2006; 188: 759-772.

21. Filliol I, Driscoll JR, Van Soolingen D, Kreiswirth BN, Kremer K, Valétudie G, Anh DD, Barlow R, Banerjee D, Bifani PJ, Brudey K, Cataldi A, Cooksey RC, Cousins DV, Dale JW, Dellagostin OA, Drobniewski F, Engelmann G, Ferdinand S, Gascoyne-Binzi D, Gordon M, Gutierrez MC, Haas WH, Heersma H, Källenius G, Kassa-Kelembho E, Koivula T, Ly HM, Makristathis A, Mammina C, Martin G, Moström P, Mokrousov I, Narbonne V, Narvskaya O, Nastasi A, Niobe-Eyangoh SN, Pape JW, Rasolofo-Razanamparany V, Ridell M, Rossetti ML, Stauffer F, Suffys PN, Takiff H, Texier-Maugein J, Vincent V, De Waard JH, Sola C, Rastogi N. 2002. Global distribution of *Mycobacterium tuberculosis* spoligotypes. *Emerg Infect Dis*. 2002 Nov; 8(11): 1347-9.

22. Filliol I, Driscoll JR, van Soolingen D, Kreiswirth BN, Kremer K, Valétudie G, Dang DA, Barlow R, Banerjee D, Bifani PJ, Brudey K, Cataldi A, Cooksey RC, Cousins DV, Dale JW, Dellagostin OA, Drobniewski F, Engelmann G, Ferdinand S, Gascoyne-Binzi D, Gordon M, Gutierrez MC, Haas WH, Heersma H, Kassa-Kelembho E, Ho ML, Makristathis A, Mammina C, Martin G, Moström P, Mokrousov I, Narbonne V, Narvskaya O, Nastasi A, Niobe-Eyangoh SN, Pape JW, Rasolofo-Razanamparany V, Ridell M, Rossetti ML, Stauffer F, Suffys PN, Takiff H, Texier-Maugein J, Vincent V, de Waard JH, Sola C, Rastogi N. 2003. Snapshot of moving and expanding clones of *Mycobacterium tuberculosis* and their global distribution assessed by spoligotyping in an international study. *J Clin Microbiol*. 41(5): 1963-1970.

23. Gagneux S, DeRiemer K, Van T, Kato-Maeda M, de Jong BC, Narayanan S, Nicol M, Niemann S, Kremer K, Gutierrez MC, Hilty M, Hopewell PC, Small PM. 2006. Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *Proc Natl Acad Sci USA*. 103(8): 2869-73.

24. García de Viedma D, Mokrousov I, Rastogi N. 2011. Innovations in the molecular epidemiology of tuberculosis. *Enferm Infecc Microbiol* Clin. 29 (Suppl 1): 8-13.

25. Glynn JR, Whiteley J, Bifani PJ, Kremer K, van Soolingen D. (2002). Worldwide occurrence of Beijing/W strains of *Mycobacterium tuberculosis*: a systematic review. *Emerg Infect Dis*. 8(8): 843-849.

26. Gordon SV, Brosch R, Billault A, Garnier T, Eiglmeier K, Cole ST. 1999. Identification of variable regions in the genomes of tubercle bacilli using bacterial artificial chromosome arrays. *Mol Microbiol*. 32: 643-655.

27. Groenheit R, Ghebremichael S, Svensson J, Rabna P, Colombatti R, Riccardi F, Couvin D, Hill V, Rastogi N, Koivula T, Källenius G. 2011. The Guinea-Bissau family of *Mycobacterium tuberculosis* complex

Réseaux

revisited. *PLoS One*. 6(4): e18601.

28. Groenheit R, Ghebremichael S, Pennhag A, Jonsson J, Hoffner S, Couvin D, Koivula T, Rastogi N, Källenius G. 2012. *Mycobacterium tuberculosis* strains potentially involved in the TB epidemic in Sweden a century ago. *PLoS One*. 7(10): e46848.

29. Gutacker MM, Mathema B, Soini H, Shashkina E, Kreiswirth BN, Graviss EA, Musser JM. Single-nucleotide polymorphism-based population genetic analysis of *Mycobacterium tuberculosis* strains from 4 geographic sites. *J Infect Dis*. 2006; 193: 121-128.

30. Heersma HF, Kremer K, van Embden JD. 1998. Computer analysis of IS6110 RFLP patterns of *Mycobacterium tuberculosis. Methods Mol Biol.* 101: 395-422.

31. Hermans PW, van Soolingen D, Bik EM, de Haas PE, Dale JW, van Embden JD. 1991. Insertion element IS987 from *Mycobacterium bovis* BCG is located in a hot-spot integration region for insertion elements in *Mycobacterium tuberculosis* complex strains. *Infect Immun*. 59: 2695-2705.

32. Hershberg R, Lipatov M, Small PM, Sheffer H, Niemann S, Homolka S, Roach JC, Kremer K, Petrov DA, Feldman MW, Gagneux S. 2008. High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. *PLoS Biol*. 2008, 6(12): e311.

33. Hirsh AE, Tsolaki AG, DeRiemer K, Feldman MW, Small PM. Stable association between strains of *Mycobacterium tuberculosis* and their human host populations. *Proc Natl Acad Sci USA*. 2004; 101: 4871-4876.

34. Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, Hill DP, Kania R, Schaeffer M, St Pierre S, Twigger S, White O, Rhee SY. (2008) Big data: The future of biocuration. *Nature*. 455(7209): 47-50.

35. Jagielski T, van Ingen J, Rastogi N, Dziadek J, Mazur PK, Bielecki J. 2014. Current Methods in the Molecular Typing of Mycobacterium tuberculosis and Other Mycobacteria. *Biomed Res Int*. 2014 (Article ID 645802): 1-21.

36. Joshi KR, Dhiman H, Scaria V. (2013) tbvar: a comprehensive genome variation resource for *Mycobacterium tuberculosis*. Database Vol. 2013: article ID bat083.

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

38. Kato-Maeda M, Rhee JT, Gingeras TR, Salamon H, Drenkow J, Smittipat N, Small PM. 2001. Comparing genomes within the species *Mycobacterium tuberculosis*. *Genome Res*. 11: 547-554.

39. Kisa O, Tarhan G, Gunal S, Albay A, Durmaz R, Saribas Z, Zozio T, Alp A, Ceyhan I, Tombak A, Rastogi N. (2012). Distribution of spoligotyping defined genotypic lineages among drug-resistant *Mycobacterium tuberculosis* complex clinical isolates in Ankara, Turkey. *PLoS One*. 2012; 7(1): e30331.

40. Koro Koro F, Kamdem Simo Y, Piam FF, Noeske J, Gutierrez C, Kuaban C, Eyangoh SI. 2013. Population dynamics of *tuberculous Bacilli* in Cameroon as assessed by spoligotyping. *J Clin Microbiol.* 51: 299-302.

41. Mahairas GG, Sabo PJ, Hickey MJ, Singh DC, Stover CK. 1996. Molecular analysis of genetic differences between *Mycobacterium bovis* BCG and virulent M. bovis. J Bacteriol. 178: 1274-1282.

42. Mokrousov I. 2012. The quiet and controversial: Ural family of *Mycobacterium tuberculosis. Infect Genet Evol.* 12: 619-629.

43. Mustafa Ali R, Trovato A, Couvin D, Al-Thwani AN, Borroni E, Dhaer FH, Rastogi N, Cirillo DM. 2014. Molecular Epidemiology and Genotyping of *Mycobacterium tuberculosis* Isolated in Baghdad. BioMed Research International, vol. 2014, Article ID 580981, 15 pages.

44. Parwati I, van Crevel R, van Soolingen D. 2010. Possible underlying mechanisms for successful emergence of the *Mycobacterium tuberculosis* Beijing genotype strains. *Lancet Infect Dis*. 10: 103-111.

45. Radhakrishnan I, K MY, Kumar RA, Mundayoor S. 2001. Implications of low frequency of IS6110 in fingerprinting field isolates of *Mycobacterium tuberculosis* from Kerala, India. *J Clin Microbiol*. 39: 1683.

46. Rastogi N, Sola C. 2007. Chapter 2 – Molecular evolution of the *Mycobacterium tuberculosis* complex. In Tuberculosis 2007: from basic science to patient care, Edited by Palomino JC, Leao S, Ritacco V. 2007, 53-91, Amedeo Online Textbooks; http://pdf.flyingpublisher.com/tuberculosis2007.pdf

47. Reyes JF, Francis AR, Tanaka MM. 2008. Models of deletion for visualizing bacterial variation: an application to tuberculosis spoligotypes. *BMC Bioinformatics*. 9: 496.

48. Rodriguez-Campos S, González S, de Juan L, Romero B, Bezos J, Casal C, Álvarez J, Fernández-de-Mera IG, Castellanos E, Mateos A, Sáez-Llorente JL, Domínguez L, Aranaz A; Spanish Network on Surveillance Monitoring of Animal Tuberculosis. 2012. A database for animal tuberculosis (mycoDB.es) within the context of the Spanish national program for eradication of bovine tuberculosis. *Infect Genet Evol.* 12(4): 877-82.

49. Rothschild BM, Martin LD, Lev G, Bercovier H, Bar-Gal GK, Greenblatt C, Donoghue H, Spigelman M, Brittain D. 2001. *Mycobacterium tuberculosis* complex DNA from an extinct bison dated 17,000 years before the present. *Clin Infect Dis*. 33: 305-311.

50. Salamon H, Segal MR, Ponce de Leon A, Small PM. 1998. Accommodating error analysis in comparison and clustering of molecular fingerprints. *Emerg Infect Dis*. 4: 159-168.

51. Schneider TD, Stephens RM. 1990. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 18: 6097-6100.

52. Shabbeer A, Cowan LS, Ozcaglar C, Rastogi N, Vandenberg SL, Yener B, Bennett KP. 2012. TB-Lineage: an online tool for classification and analysis of strains of *Mycobacterium tuberculosis* complex. *Infect Genet Evol.* 12: 789-797.

53. Sheen P, Couvin D, Grandjean L, Zimic M, Dominguez M, Luna G, Gilman RH, Rastogi N, Moore DA. 2013. Genetic diversity of *Mycobacterium tuberculosis* in Peru and exploration of phylogenetic associations with drug resistance. *PLoS One*. 8(6): e65873.

54. Smit PW, Haanperä M, Rantala P, Couvin D, Lyytikäinen O, Rastogi N, Ruutu P, Soini H. (2013). Molecular epidemiology of tuberculosis in Finland, 2008-2011. *PLoS One*. 8(12): e85027.

55. Smith NH, Upton P. 2012. Naming spoligotype patterns for the RD9-deleted lineage of the *Mycobacterium tuberculosis* complex; www.Mbovis.org. *Infect Genet Evol*. 12 (4), pp. 873–876.

56. Soares P, Alves RJ, Abecasis AB, Penha-Gonçalves C, Gomes MG, Pereira-Leal JB. (2013) inTB – a data integration platform for molecular and clinical epidemiological analysis of tuberculosis. *BMC Bioinformatics*. 14: 264.

57. Sola C, Devallois A, Horgen L, Maïsetti J, Filliol I, Legrand E, Rastogi N. 1999. Tuberculosis in the Caribbean: using spacer oligonucleotide typing to understand strain origin and transmission. *Emerg Infect Dis*. 5(3): 404-14.

58. Sola C, Filliol I, Gutierrez MC, Mokrousov I, Vincent V, Rastogi N. 2001. Spoligotype database of *Mycobacterium tuberculosis*: biogeographic distribution of shared types and epidemiologic and phylogenetic perspectives. *Emerg Infect Dis*. 7(3): 390-396.

59. Sola C, Filliol I, Legrand E, Mokrousov I, Rastogi N. 2001. *Mycobacterium tuberculosis* phylogeny reconstruction based on combined numerical analysis with IS1081, IS6110, VNTR, and DR-based spoligotyping suggests the existence of two new phylogeographical clades. *J Mol Evol*. 53: 680-689.

60. Sola C, Filliol I, Legrand E, Lesjean S, Locht C, Supply P, Rastogi N. 2003. Genotyping of the *Mycobacterium tuberculosis* complex using MIRUs: association with VNTR and spoligotyping for molecular epidemiology and evolutionary genetics. *Infect Genet Evol*. 3: 125-133.

61. van Soolingen D, Qian L, de Haas PE, Douglas JT, Traore H, Portaels F, Qing HZ, Enkhsaikan D, Nymadawa P, van Embden JD. (1995). Predominance of a single genotype of *Mycobacterium tuberculosis* in countries of East Asia. *J Clin Microbiol.* 33: 3234-3238.

62. Sreevatsan S, Pan X, Stockbauer KE, Connell ND, Kreiswirth BN, Whittam TS, Musser JM. 1997. Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. *Proc Natl Acad Sci USA*. 94: 9869-9874.

Réseaux

63. Supply P, Lesjean S, Savine E, Kremer K, van Soolingen D, Locht C. 2001. Automated high-throughput genotyping for study of global epidemiology of *Mycobacterium tuberculosis* based on mycobacterial interspersed repetitive units. *J Clin Microbiol*. 39: 3563-3571.

64. Supply P, Allix C, Lesjean S, Cardoso-Oelemann M, Rüsch-Gerdes S, Willery E, Savine E, de Haas P, van Deutekom H, Roring S, Bifani P, Kurepina N, Kreiswirth B, Sola C, Rastogi N, Vatin V, Gutierrez MC, Fauville M, Niemann S, Skuce R, Kremer K, Locht C, van Soolingen D. 2006. Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis. J Clin Microbiol*. 44: 4498-4510.

65. Tang C, Reyes JF, Luciani F, Francis AR, Tanaka MM. 2008. SpolTools: online utilities for analyzing spoligotypes of the *Mycobacterium tuberculosis* complex. *Bioinformatics.* 24: 2414-415.

66. Wolfe ND, Dunavan CP, Diamond J. 2007. Origins of major human infectious diseases. *Nature*. 447: 279-283.

67. Zink AR, Sola C, Reischl U, Grabner W, Rastogi N, Wolf H, Nerlich AG. 2003. Characterization of *Mycobacterium tuberculosis* complex DNAs from Egyptian mummies by spoligotyping. *J Clin Microbiol*. 41: 359-367.

68.Weniger T, Krawczyk J, Supply P, Niemann S, Harmsen D. 2010. MIRU-VNTRplus: a web tool for polyphasic genotyping of *Mycobacterium tuberculosis* complex bacteria. *Nucleic Acids Res*. 38(Web Server issue): W326–W331.

69.WHO. 2006. The Stop TB Strategy: Building on and enhancing DOTS to meet the TB-related Millennium Development Goals. http://whqlibdoc.who.int/hq/2006/WHO_HTM_STB_2006.368_eng.pdf?ua=1.

70.WHO. 2013. Global tuberculosis report 2013.

www.who.int/iris/bitstream/10665/91355/1/9789241564656_eng.pdf