Etat de l'art du projet de recherche

Stephane Robin

1er janvier 2020

1. Introduction

Le présent document cherche à définir les fondements sur lesquels va reposer le stage et en particulier de connaître l'état actuel des connaissances concernant l'analyse bioinformatique du locus CRISPR-Cas chez *Mycobacterium tuberculosis*.

L'objectif du stage est de comprendre les liens entre les sept lignées de *tuberculosis* et les alternances de gènes et espaceurs du locus CRISPR-Cas, au moyen d'outils développés au sein de l'équipe AND de l'Université de Franche Comté.

2. Evolution de la branche humaine de la tuberculose

Le développement des maladies s'adapte à la densité de population concernée. En effet, auprès d'une foule dense, les infections se répandent plus largement et deviennent plus virulentes, alors qu'auprès d'une population moins importante, elles ont une croissance plus faible, laissant place parfois à des périodes où les infections restent latentes.

Une période charnière dans l'histoire de l'humanité est la transition démographique du Néolithique, qui a vu il y a 10 000 ans, suite à l'apparition de l'agriculture et de l'élevage, un accroissement de la population, favorisant la naissance de nombreuses maladies. Les maladies humaines plus anciennes se développaient auprès de populations moins denses et produisaient des phases chroniques de latence et de réactivation permettant aux populations infectées de survivre.

La tuberculose conjugue ces deux modèles de maladie. En effet, elle a montré à travers les âges des périodes de réactivation, elle dépend fortement de la densité de population et son mode de transmission aérosol s'est parfaitement adapté aux foules.

L'analyse phylogénique [1] de Comas et al.[1] se base exclusivement sur l'étude du génome [2] complet de toutes les lignées connues de tuberculose en utilisant les SNPs [3] pour construire les relations entre les différentes branches. Les résultats obtenus rejoignent de précédentes études effectuées à partir d'autres marqueurs, et confirment l'existence de sept principales lignées de tuberculose. On remarque en particulier que plusieurs branches d'origine animale se sont regroupées avec la lignée 6 d'Afrique de L'Ouest 2, et que les lignées modernes 2, 3 et 4 d'Europe ont des origines proches. Par ailleurs, seuls 1,1 % des SNPs sont homoplastiques [4], ce qui suggère que la structure de la tuberculose favorise les clonages plutôt que les recombinaisons entre branches.

L'étude phylogénique de Comas et al.[1] corrobore les connaissances actuelles selon lesquelles la tuberculose est originaire d'Afrique. Par ailleurs, s'appuyant sur les origines africaines de l'espèce humaine, elle cherche également à déterminer l'ancienneté de l'association entre la tuberculose et

1. **Phylogénie** : étude des liens entre espèces apparentées, permettant de retracer les principales étapes de l'évolution des organismes depuis un ancêtre commun.
2. **Génome** : ensemble de l'information génétique d'un organisme. Par extension, le génome se réfère aussi au support physique de cette information génétique, la macromolécule d'ADN.
Annotation des gènes : processus permettant d'identifier l'emplacement des gènes dans l'ADN, de déterminer leurs fonctions et leurs possibles interactions.
3. **SNP Single Nucleotide Polymorphism** : il s'agit de la variation ou polymorphisme d'une seule paire de bases du génome entre organismes d'une même espèce. Les SNPs sont peu sujettes à des phénomènes d'homoplasie et sont de ce fait utilisées pour des études phylogénétiques et de classification
4. **Homoplasie** : similitude de caractères chez différentes espèces, qui ne provient pas d'un ancêtre commun, mais peut par exemple provenir d'une adaptation à l'environnement. Diffère de l'homologie qui est une similitude de caractères observée chez deux espères différentes, provenant de l'héritage d'un ancêtre commun.

son hôte humain. L'analyse des divergences des génomes de la tuberculose est comparée à celle d'une arborescence génétique déjà établie à partir de mitochondries [5] de l'être humain.

Les similitude relevées montrent que la tuberculose a infecté les premiers hommes d'Afrique. Pour aller plus loin, l'étude de Comas et al. a tenu compte de 3 dates importantes dans l'évolution de l'être humain qui ont été reportées sur l'analyse phylogénique de la tuberculose des lignées 5 et 6 d'Afrique de l'Ouest :

- l'émergence de l'homo sapiens correspondant au MTBC-185 [6],
- l'émergence du haplogroupe [7] mitochondriaque de la lignée 3 chez l'homme correspondant au MTBC-70,
- le début de la transition démographique du Néolithique correspondant au MTBC-10.

La branche MTBC-70 a révélé des corrélations avec l'histoire de l'humanité, telle qu'elle a été décrite par l'archéologie, en montrant l'apparition des sept différentes lignées de tuberculose :

- il y a 73 000 ans, apparition des lignées 5 et 6 correspondant à une 1ère migration humaine importante vers l'Afrique de l'Ouest,
- il y a 67 000 ans, apparition de la lignée 1 correspondant à une migration humaine importante autour de l'Océan Indient,
- il y a 64 000 ans, apparition de la lignée 7 concernant une population qui est restée en Afrique ou est revenue en Afrique après une 1ère migration,
- il y a 46 000 ans, apparition de la lignée 4 correspondant à une migration humaine importante vers l'Europe,
- il y a 42 000 ans, apparition des lignées 2 et 3 correspondant à une migration humaine importante vers l'Asie de l'Est et l'Asie Centrale.

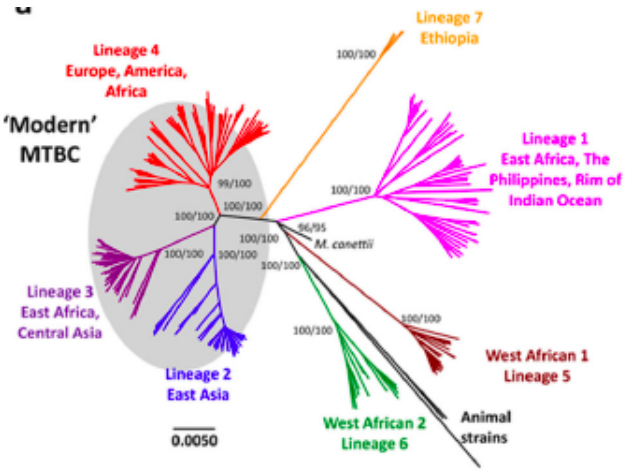


Figure 2.1 – Philogénie du génome complet de MTBC, d'après *Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans*

En revanche, la branche MTBC-185 suggère l'apparition de mutations à partir de lignées africaines il y a 174 000 ans, c'est à dire que la dispersion de la tuberculose précèderait celle de l'homo sapiens.

Dans tous les cas, la tuberculose aurait infecté l'espèce humaine et évolué conjointement avec elle depuis 70 000 ans, mais son apparition serait antérieure à la transition démographique du Néolithique.

5. **Mitochondrie** : centrale énergétique des cellules qui contribue à la production d'ATP.
6. **MTBC** : Mycobacterium Tuberculosis Complex.
7. **Haplogroupe** : groupe possédant les mêmes caractères génétiques et partageant un ancêtre commun suivant une mutation SNP.

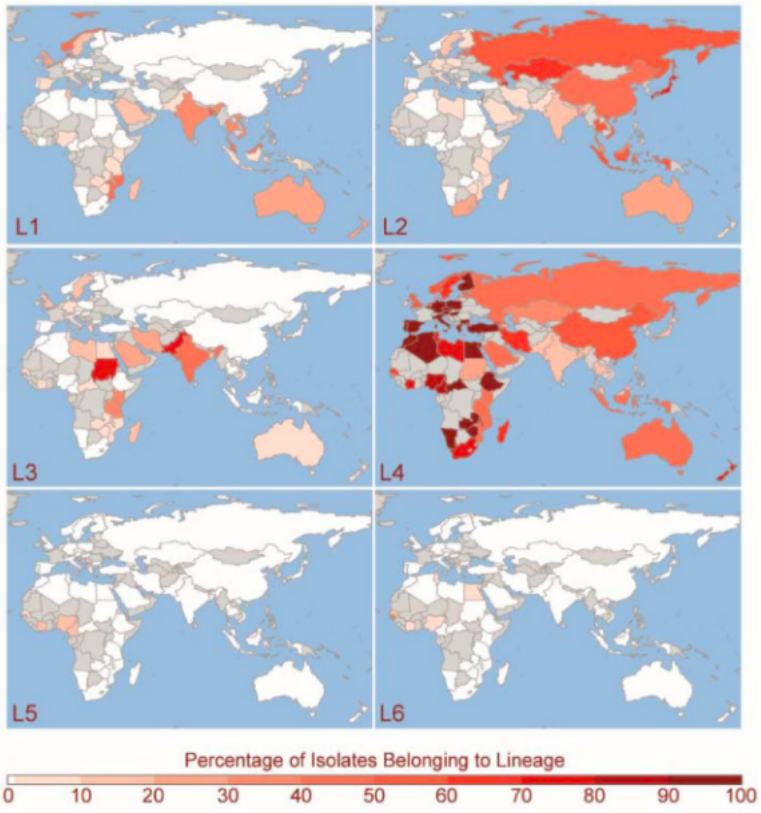


Figure 2.2 – Distribution géographique des lignées 1 à 6, d'après
Lineage specific histories of Mycobacterium tuberculosis dispersal in Africa and Eurasia

La base de données de tuberculose étudiée de façon probabiliste par Comas et al. montre que le Néolithique a fortement contribué à l'expansion de la maladie il y a 10 000 ans grâce à l'augmentation de la densité de population, à la probabilité de co-infection avec d'autres maladies dépendantes de la densité de population, et non pas grâce à la possibilité pour la tuberculose de muter d'une variété animale vers une variété humaine. En effet, l'analyse phylogénique de la tuberculose montre que les branches humaines ont divergé des branches animales avant le Néolithique.

Le Néolithique n'était pas la seule période où l'augmentation de la population fut importante, toutefois la concentration de population qui s'en est suivie a permis l'apparition auprès de la tuberculose de caractères fortement dépendants de la densité de population qu'elle affecte. Le Néolithique a donc marqué un tournant dans l'histoire de la tuberculose qui a alors commencé à conjuguer les deux modèles de maladie, dépendant de la densité de population et chronique par périodes de réactivation.

Il faut donc considérer que la co-existence de la tuberculose avec l'espèce humaine depuis des milliers d'années a conduit la maladie à s'adapter aux changements du génome humain et inversement. Les prochaines études sur la tuberculose devraient donc se concentrer sur des génomes complets de la tuberculose et de l'être humain choisis en rapport à leurs associations.

Par ailleurs, une étude récente de Perry S. et al.[7, 8] suggère que l'infection d'une organisme par l'Helicobacter Pylori pourrait protéger de la tuberculose sous sa forme active. A contrario, nous ne savons pas si la tuberculose latente pourrait protéger contre les ulcers et les cancer de l'estomac causés par l'Helicobacter Pylori.

3. L'expansion de la lignée 4 de tuberculosis

La lignée 4 de tuberculosis est la plus répandue de par le monde. Dans leur article [3], Brynildsrud O.B. et al. montrent que la dispersion de la lignée 4 de tuberculosis est liée à l'expansion coloniale européenne en Afrique et en Amérique entre le 17è et le 19è siècle.

Brynildsrud O.B. et al. utilisent des méthodes d'analyse discrète et une approche bayésienne en phylogénie moléculaire [8] pour obtenir de manière formelle l'évolution phylogéographique de la lignée 4 de tuberculosis. Ils estiment que le plus récent ancêtre commun MRCA de la lignée 4 est apparu en Europe en 1096 après JC en Europe. Cela ne contradit pas les résultats de O'Neil M.B. et al. qui estiment dans [14] l'origine de la lignée 4 autour de la méditerranée, si l'on considère l'Europe en tant que continent dans un sens large.

L'analyse phylogeographique de Brynildsrud O.B. suggère que les premières vagues de migration de la lignée 4 hors d'Europe se sont déroulées au début du 13ème siècle vers l'Asie du Sud Est. En particulier, la structure des isolats Vietnamiens correspond bien avec celle des souches connues à l'époque de l'expansion coloniale française en Indochine.

Les vagues suivantes de migration de la lignée 4 se sont dirigées vers l'Afrique de l'Ouest au 15ème siècle, puis en Afrique de l'Est et du Sud au 17ème siècle. Les échanges continus avec le Portugal au 15ème siècle ont favorisé la dispersion de la maladie, ce qui a été renforcé plus tard par la colonisation française de l'Afrique de l'Ouest Ces échanges avec les populations européennes ont prévalu à une transmission locale de la tuberculose jusqu'au 19ème siècle.

La première migration interne de la maladie en Afrique date de l'Empire Zulu au 19ème siècle et se dirigeait vers le Nord et l'Est africain.

La transmission de la maladie en Amérique date quant à elle du 15ème siècle avec la colonisation du continent, mais il faudra attendre le 17ème siècle pour voir l'explosion de la maladie en Amérique du Sud. Ce retard dans l'évolution de la maladie par rapport à la branche africaine peut s'expliquer par le taux de mortalité élevé des populations aborigènes au contact des européens.

émergence de la résistance antimicrobienne AMR antimicrobal resistance with multidrug-resistant tb (MDR-TB) in recent decades is overwhelmingly a local phenomenon, while migration of resistant strains seems to have played a marginal role in shaping the observed L4 AMR landscape

à l'âge des antibiotiques, des mutations conférant des antimicrobiens une résistance écrasante a émergé localement (au niveau des nations), avec une transmission transfrontalière minimale de résistance.

MTBC est considéré comme étant résistant aux antibiotiques.

4. Le locus CRISPR-Cas

4.1. Quelques caractéristiques du génome de M. tuberculosis

La souche H37Rv de M. tuberculosis est la souche de tuberculose la plus étudiée en laboratoire, depuis sa découverte en 1905. Elle sert aujourd'hui de référence pour le séquençage et l'annotation du génome de M. tuberculosis. Constitué d'environ 4 millions de paires de base et 4000 gènes, il se caractérise par un taux élevé de guanine G et de cytosine C (65,6%), et un codon GTG qui sert de

8. **Approche bayésienne** : méthode probabiliste basée sur le calcul des probabilités postérieures des arbres phylogénétiques par la combinaison d'une probabilité antérieure avec la fonction de vraisemblance

codon d'initiation dans 35% des gènes. L'homologie entre les ADN des différentes sous-espèces de M. tuberculosis est très élevée ce qui pourrait peut-être s'expliquer par un tax de mutation très faible.

Parmi les marqueurs génétiques utilisés pour des études phylogéniques ou d'épidémiologie moléculaire, on retrouve les SNPs, le locus CRISPR, les MIRU [9], et les VNTR [10]. L'association des résultats obtenus par ces marqueurs génère un profil allélique utile pour l'étude du complexe M tuberculosis. La base de données SITVIT présentée dans l'article[12] contient les génotypes de tuberculosis obtenus à partir des trois marqueurs moléculaires : spoligotypes, MIRU et VNTR.

4.2. Description du locus CRISPR-Cas

Le locus CRISPR *Clusterd Regularly Interspaced Short Palindromic Repeats* est une famille de séquences répétées (DR pour *Direct Repeat*) dans l'ADN formant un palindrome, qui se trouve à l'état naturel chez 40% des bactéries (dont le MTBC) et la plupart des archées. Elle est héritable par transmission aux cellules filles et se conserve donc pour une même espèce. Chez M. tuberculosis chaque série de répétition contient 36 bp [11], régulièrement espacées par des espaceurs de 34 à 41 bp. 104 espaceurs ont été retrouvés jusqu'à maintenant dans toutes les souches du M. tuberculosis. Les locus CRISPR sont généralement adjacents aux gènes Cas, dont ils sont séparés par une séquence de 300 à 500 bp, appelée leader qui contrôle à la fois la l'acquisition de l'ADN viral par les spacers et la fabrication de protéines. Les gènes Cas produisent des protéines aux fonctionnalités différentes et notamment les enzymes [12] capables de couper l'ADN en vue de leur réparation.

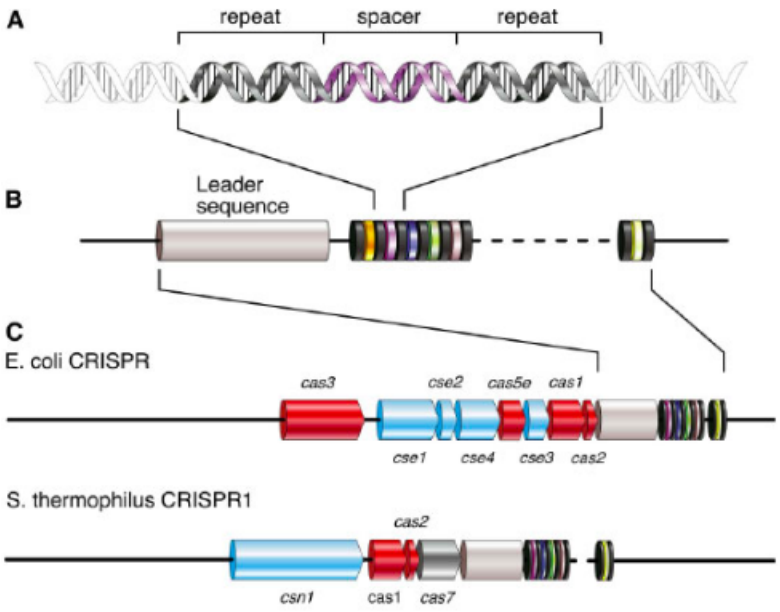


Figure 4.3 – Structure du locus CRISPR-Cas, d'après
https://www.sinobiological.com/crispr-locus.html

9. **marqueurs MIRU** : séquences nucléotiques courtes répétitives en tandem entrecoupées de mycobactéries. La méthode MIRU actuellement utilisée sur Mycobacterium tuberculosis est composée de 12 loci MIRU différents. Un mirutype est un modèle à 12 chiffres représentant le nombre de répétitions de chacun de ces 12 loci spécifiques.
10. **marqueures VNTR** : séquences nucléotiques courtes en tandem à nombre variable. Cinq répétitions en tandem exactes (locus ETR) sont utilisées pour l'analyse VNTR du complexe M. tuberculosis.
11. **bp** : une paire de base est l'appariement de 2 bases nucléiques situées sur 2 brins complémentaires d'ADN, reliées par des ponts d'hydrogène.
12. **Enzyme de restriction** : protéine capable de couper un fragment d'ADN au niveau d'une séquence de nucléotides caractéristique appelée site de restriction. Chaque enzyme de restriction reconnaît ainsi un site spécifique.

Ces séquences incorporent dans les spacers des fragments d'ADN de bactériophages qui ont déjà infecté la bactérie qui sont stockés en vue de détecter et de détruire l'ADN de bactériophages similaires en cas d'infection ultérieure. Par conséquent, CRISPR-Cas est un système immunitaire naturel utilisé par les bactéries pour se protéger des infections virales.

4.3. Fonctionnement du système CRISPR-Cas

Les systèmes CRISPR-Cas sont de trois types et utilisent les différents gènes Cas pour intégrer des fragments de gènes étrangers dans les spacers de CRISPR. Par exemple, dans le cas d'une bactérie qui détecte la présence d'ADN ou d'ARN d'un virus, elle produit une enzyme nucléase appelée Cas9 capable de couper l'ADN viral, puis une séquence d'ARN CRISPR notée crARN correspondant à celle de l'ADN du virus et servant de guide ARN, puis finalement une séquence d'ARN traceur notée trARN. Lorsque trARN trouve sa cible parmi le génome du virus, Cas9 sectionne l'ADN viral puis en incorpore un fragment dans un *spacer* du génome de la bactérie, conservant ainsi en mémoire une trace de ce virus en vue d'une éventuelle infection future. Les espaceurs servent donc de banque de mémoire en conservant l'ADN des virus qui ont attaqué la bactérie.

La technologie CRISPR-Cas9 s'inspire du système du même nom a d'abord été utilisée pour typer les souches bactériennes, suivant une technique appelée *spoligotyping*. CRISPR-Cas9 est actuellement principalement employé comme ciseau moléculaire afin d'éditer le génome et d'introduire localement des modifications génétiques.

4.4. Le spoligotypage

La région DR du locus CRISPR-Cas présente un niveau de polymorphisme suffisant pour pouvoir classer phylogéographiquement les souches du MTBC. Le polymorphisme entre les différentes souches résulte des variations et de l'identité des spacers. C'est ce polymorphisme qui est exploité en 1997 par Kamerbeek et al. et expliqué dans [16] comme technique de génotypage spécifique du MTBC. Le *Spacer Oligonucleotide Typing*, repose sur la détection de séquences répétitives trouvées entre les gènes d'un agent infectieux au sein d'un locus CRISPR-Cas. Pour ce faire, la région DR d'un isolat à tester subit un traitement par amplification PCR [13] ou celui d'une puce à ADN [15] pour dévoiler un motif de taches correspondant aux spacers.

La comparaison de ces motifs permet la différentiation des souches. Quarante-trois espaceurs les plus polymorphes ont été utilisés pour le typage des mycobactéries suivant Kamerbeek et al. La classification classique de MTBC utilise donc un groupe de 43 bit représentant la présence ou l'absence d'espaceurs dans le locus CRISPR, qu'on appelle spoligotype. Des études pour augmenter le niveau de discrimination du spoligotypage ont été faites en 2010 utilisant 68 espaceurs. A l'heure actuelle l'équipe AND de l'université de Franche Comté utilise 98 espaceurs pour ce génotypage.

La technique ne nécessite pas une importante quantité d'ADN car elle est basée sur une amplification de la région DR par PCR. Les spoligotypes ainsi obtenus peuvent être partagées entre laboratoires et corroborent les résultats obtenus à partir d'autres marqueurs génétiques. Ces données numériques

13. **PCR Polymerase Chain Reaction** : méthode de réaction en chaîne utilisant un polymère [14] pour dupliquer en grand nombre une séquence d'ADN spécifique. La méthode PCR repose sur le cycle thermique, qui expose les séquences à des cycles répétés de chauffage et de refroidissement pour permettre différentes réactions dépendantes de la température comme la fusion de l'ADN et la réplication de l'ADN par les enzymes. La méthode PCR utilise deux agents principaux : les polymères d'ADN et les *primers*.

15. **Puce à ADN** : ensemble de molécules d'ADN fixées sur une petite surface solide permettant de mesurer le niveau d'expression d'un grand nombre de gènes simultanément, ou de déterminer le génotype de plusieurs régions d'un génome.

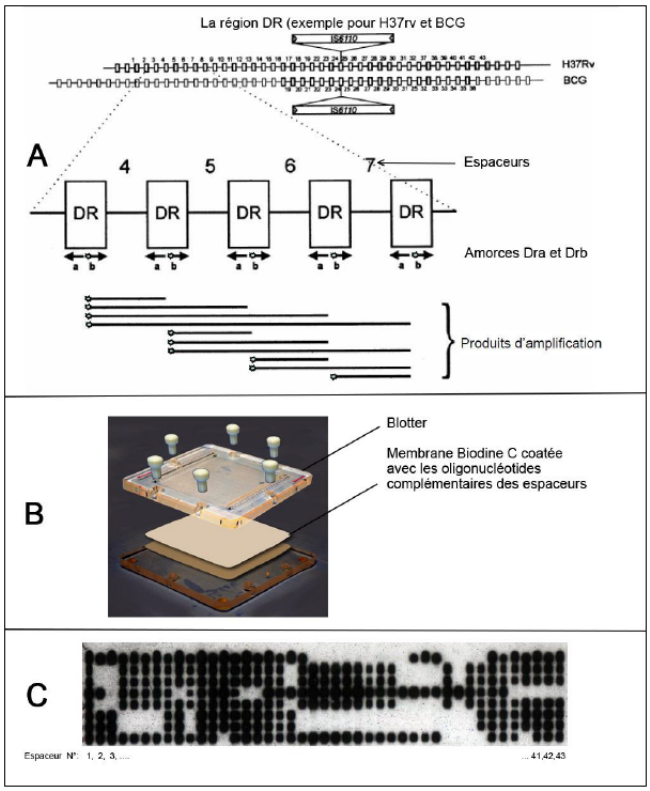


Figure 4.4 – Les différentes étapes du spoligotypage d'après *Etudes descriptive, épidémiologique, moléculaire et spatiale des souches Mycobacterium tuberculosis circulant à Antananarivo, Madagascar*

permettent de bien différencier les souches de tuberculosis et sont de moindre coût comparativement à d'autres méthodes. Cependant, le spoligotyping éprouve des difficultés à bien différencier les souches au sein de grandes familles de souches telles que la branche Beijing.

Le spoligotyping a permis de fournir une image globale de la diversité des souches de tuberculosis.

Notons qu'une nouvelle technologie Spoligo-RIF, TB-SPRINT et TB-SPRINTplus ont été développés pour combiner le spoligotypage avec des tests moléculaires de sensibilité aux antituberculeux.

4.5. Vers une normalisation des spoligotypes

Au début du spoligotyping, il n'existait pas de norme pour décrire les motifs formés par les spacers ou simplement les numéroter. Chaque laboratoire utilisait son propre système de numérotation accompagné d'un schéma descriptif du motif. Ce manque de normalisation entravait les possibilités de comparaison des résultats obtenus et le développement d'une vision mondiale de l'évolution de MTBC. Une méthode standardisée de description des spoligotypes a été proposée en 2001 par Dale JW dans son article [10].

Tout d'abord, une base de données centralisée regroupant tous les motifs connus et de leurs numérotations jusqu'à présent associées existe au RIVM Rijksinstituut voor Volksgezondheid en Milieuhygiene, Bilthoven, Netherlands. Elle peut être consultée au `http://www.caontb.rivm.nl`. A partir de 2001, les nouveaux motifs devraient prendre un unique format de numérotation pour être répertoriés dans cette base de données. Toutefois, cela nécessite l'interrogation systématique de la base de données et la comparaison avec les éléments déjà existants pour chaque nouveau spoligotype. Pour éviter cette perte de temps, de nombreux laboratoires utilisent des systèmes rationnels avec

Numéro	Numéro espaceur en spoligotypage	Numéro espaceur dans le génome	séquence des oligonucléotides (5'→3')	Numéro	Numéro espaceur en spoligotypage	Numéro espaceur dans le génome	séquence des oligonucléotides (5'→3')
1	1	2	CGCTCCCCTAGTCGT	53	-	17	TCTTGAGCAACGCCATC
2	2	3	TGGGCGACAGCTTTTGA	54	-	45	AAGTTGGCGCTGGGG
3	3	4	CTTCCAGTGATCGCCTT	55	-	48	AACCGTCCCACCTG
4	4	12	TCATACGCCGACCAATC	56	-	49	AACACTTTTTTTGAGCGTGG
5	5	13	TTCTGACCACTTGTGCG	57	-	50	CGGAAACGCAGCACC
6	6	14	TCATTTCCGGCTT	58	-	54	CGATCATGAGAGTTGCG
7	7	15	TGAGGAGAGCGAGTACT	59	-	55	TTTTCGCTGTTGTGGTTCT
8	8	18	TGAAACCGCCCCCAG	60	-	56e	AGCACCTCCCTTGACAA
9	9	19	ACTCGGAATCCCATGTG	61	-	57	TGCTGACTTCGCCTGTA
10	10	20	CTCTAGTTGACTTCCGG	62	-	58	CGAGCAGCGGCATAC
11	11	21	CAGGTGAGCAACGGC	63	-	59	GCATCCACTCGTCGC
12	12	22	ATGGGATATCTGCTGCC	64	-	60	TGGTAATTGCGTCACGG
13	13	23	ATTGCCATTCCCTCTCC	65	-	61	ACCATCCGACGCAGG
14	14	24	TTTCGGTGTGATGCGGA	66	-	66	CCACGCTACTGCTCC
15	15	25	TGAATAACGCGCAGTGAAT	67	-	67	CACCGCCGATGACAG
16	16	26	TCGCACGAGTTCCCG	68	-	68	GTGTTTCGGCCGTGC
17	17	27	CCGGCAACAATCGCG	69	-	69	GTTGCATTCGTCGACTG
18	18	28	TGCAGATGGTCCGGG	70	-	70	GGCGGCGCCGAGAA
19	19	29	TTGCGCTAACTGGCTTG	71	-	71	TTCCATGACTTGACGCC
20	20	30	ATTTCCTTGACCTCGCC	72	-	72	CGATGCGGCCACTAG
21	21	31	CGATGTCGATGTCCCAA	73	-	73	GCTGACCCCATGGATG
22	22	32	ACGGCACGATTGAGACA	74	-	74	CAACAAGGTCTACGCGT
23	23	33	GTCCAGCTCGTCCGT	75	-	75	GATCAGGCGAAGGCG
24	24	34	GCCTGCTGGGTGAGA	76	-	76	ATTGCAGCGACGGGC
25	25	35	GGAGCCGATCAGCGA	77	-	77	CAACGACGCTGTATTGG
26	26	36	CTTCAGCACCACCATCA	78	-	78	AGCAGCATGGACGGTTT
27	27	37	TTCGTGATCTCTTCCCG	79	-	79	GCGGATGTGGTGGTC
28	28	38	GATCACAACACCAACTAATG	80	-	80	GTACATAGCGAGCTG
29	29	39	GAAATACAGGCTCCACG	81	-	81	GCCGCGGGTTTCGTT
30	30	40	TCTTGACGATGCGGTTG	82	-	82	GGGGCGTGTGTTCGT
31	31	41	TTCGCGTCAGACAGGTT	83	-	83	CTGGTGTGCTTATGCCT
32	32	42	ACTCCCGACCAAATAGG	84	-	84	CAAATGTTTGGACTGTGATC
33	33	43	TCGACACCGACATGAC	85	-	85	TTGTCGCGCGCCTTTTT
34	34	44	GAAGTCACCTCGCCC	86	-	86	GTTTCAGTTTTCTTGTCCC
35	35	46	AGTCCGTACGCTCGAAA	87	-	87	CTGGTTGTTGCCCGG
36	36	47	CGAAATCCAGCACCACA	88	-	88	TGTTCGGTGTTCTCCTG
37	37	51	TTTGAGCGCGAACTCGT	89	-	89	TCATGACGAGCCCGCA
38	38	52	TGGATGGCGGATGCG	90	-	90	ACACGGCCTGATCGGT
39	39	53	AAATCGGCGTGGGTAAC	91	-	91	CGGATTGTCTGGCCC
40	40	62	TCATACAGGTCCAGTGC	92	-	92	TAAGCACGCGTCTGTCA
41	41	63	GCTTTCCGGCTTCTATC	93	-	93	GACCACCGAATCACCAT
42	42	64	GACATGGACGAGCGC	94	-	94	TCTGGTAGTGGGCTTCT
43	43	65	CAGAATCGCACCGGG	95	-	11	ACATGCCGTGGCTCA
44	-	1	CAACCCGGAATTCTTGC	96	-	16	CACGACGTTAGGGCA
45	-	5	CAGGCGTGGCTAGG	97	-	5	CGGCAGGCGTGGCTA
46	-	6	GTCGCCGTAAGTGCC	98	-	6	CCGTCGCCGTAAGTG
47	-	7	GTTGACCACGAATTTTCAG	99	-	17	GAGCAACGCCATCAT
48	-	8	GCTGGCGCGCATCAT	100	-	11	TGAGCCACGGCATGT*
49	-	9	CCATATCGGGGACGG	101	-	16	ATGCCCTAACGTCGTG*
50	-	10	GCGTCGTGCCATCAG	102	-	5	TAGCCACGCCTGCCG*
51	-	11	CCGTGCACATGCCGT	103	-	6	CACTTACGGCGACGG*
52	-	16	ACGTTAGGGCATGCAG	104	-	17	GATGATGGCGTTGCT*

Figure 4.5 – Espaceurs connus chez MBTC, d'après *Etudes descriptive, épidémiologique, moléculaire et spatiale des souches Mycobacterium tuberculosis circulant à Antananarivo, Madagascar*

des codes descriptifs assignés à chaque isolat.

Dale JW et al. dans leur article [10] proposent d'utiliser exclusivement un système rationnel octal ou hexadécimal, sachant qu'il est aisé de passer de l'un à l'autre et qu'il est également facile de retrouver l'état initial binaire. Ainsi, les motifs de spoligotype comprenant 43 bits seraient réduits dans le système octal en 14 groupes de 3 bits auquel s'ajouterait un unique bit, ce qui donnerait finalement un ensemble de 15 chiffres en écriture octale. En ce qui concerne le système hexadécimal, les motifs de 43 bits seraient réduits en 6 groupes de 8 bits avec un dernier groupe ne comprenant que 3 bits, soit 6 groupes de 2 chiffres hexadécimaux. Notons que un bit symbolise dans ce cas la présence ou l'absence d'une espaceur dans le locus étudié.

Le site Mbovis.org, `https://www.mbovis.org/database.php`, bien que dédié aux souches animale

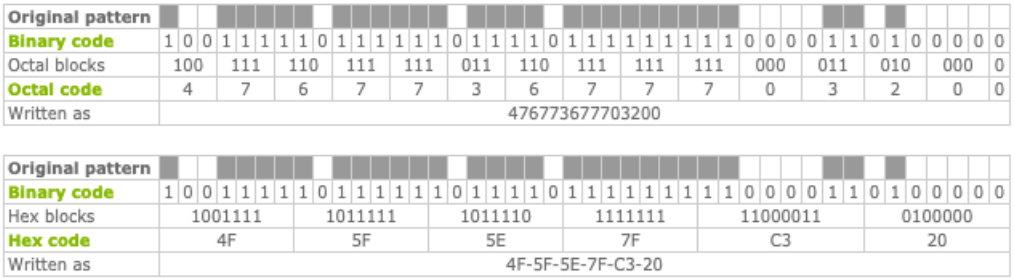


Figure 4.6 – Exemple de système rationnel octal et hexadécimal, d'après https://www.mbovis.org/spoligotype-nomenclature.php

de MTBC, fournit une application pratique permettant de transformer rapidement les spoligotypes binaires en système octal ou hexadéciamal.

4.6. Quel outil informatique pour le spoligotypage ?

Les technologies PCR de génotypage utilisent toujours différents marqueurs tels que les SNPs pour obtenir en laboratoire des résultats fiables. Des logiciels informatique de prédiction du génotype ont également fait leur apparition pour optimiser les coûts et le gain de temps. Ils offrent un outil de comparaison des résultats obtenus expérimentalement et in silico.

SpolPred est un logiciel de prédiction rapide et précis des spoligotypes de tuberculosis à partir de séquences génomiques courtes appelées reads [16]. Cet outil développé par Preston M. fonctionne efficacement avec des reads provenant de plateformes telles que Illumina GAII ou HiSeq. SpolPred utilise des fichiers de séquences de reads simples ou par paires au format FASTQ, pour produire une prédiction de spoligotype au format octal qui est comparée au spoligotype correspondant dans la base SITVITWEB [17].

Dans leur étude, Coll F. et al.[2] montrent en 2012 l'utilité de SpolPred en comparant les spoligotypes obtenus par le logiciel avec les résultats de laboratoire. Ils dévoilent ainsi les limites de la méthode expérimentale qui a répertorié cinq faux spoligotypes, alors que SpolPred a su éviter une erreur de classification du génotype. Par ailleurs, il apparaît que SpolPred offre plus de rapidité et des résultats pratiquement identiques à ceux obtenus avec la méthode bioinformatique par assemblage, consistant à fusionner des fragments d'ADN issus d'une plus longue séquence afin de reconstruire la séquence originale, à l'aide du logiciel Velvet développé en 2008.

Toutefois, d'après l'étude de Xia et al.[9], la précision de SpolPred est fortement réduite lorsque les reads n'ont pas une taille uniforme, comme par exemple lorsqu'ils proviennent de séquençages Ion Torrent ou de la plateforme de diagnostique clinique Illumina MiSeq. Ainsi, lorsque les reads ne sont pas uniformes, la précision des résultats dépend fortement de leurs tailles et donc du choix initial fait par l'opérateur. Par ailleurs, SpolPred demande à l'utilisateur de spécifier la direction de lecture des reads, et le logiciel n'utilise donc qu'une partie des informations fournies par les reads.

Une problématique de SpolPred en 2020 est que le logiciel n'est plus disponible au public. En effet, une visite sur le site officiel http://www.pathogenseq.org/spolpred fourni comme référence dans le document [2] de Coll F. et al. montre que le nom de domaine est à vendre. Preston M., qui a fait partie

16. **Read** : mélange de courtes séquences oligonocléotidiques de 20 à 200 pb généré par des séquenceurs
17. **SITVITWEB** : application en ligne connectée à la base de données SITVIT http://www.pasteur-guadeloupe.fr:8081/SITVIT_ONLINE/query et permettant d'analyser des data au regard des markers Spoligotypes, MIRU et VNTR. L'application en ligne SITVIT2 http://www.pasteur-guadeloupe.fr:8081/SITVIT2/index.jsp est une mise à jour de SITVITWEB qui est pourtant toujours en activité

de l'équipe de rechercher de Coll F. pour le développement de SpolPred, a bien créé un site https://www.mybiosoftware.com/spolpred-predict-the-spoligotype-from-raw-sequence-reads.html proposant le téléchargement du logiciel, mais le lien est actuellement inactif.

Une alternative à SpolPred est SpoTyping présenté en 2016 dans l'article [9] de Xia et al. comme étant 20 à 40 fois plus rapide que SpolPred pour prédire avec précision des spoligotypes de tuberculosis à partir de reads de taille uniforme ou variable. Par ailleurs, SpoTyping lit chaque read dans les deux directions en exploitant complètement les informations fournies. SpoTyping réduit la durée des recherches en intégrant l'algorithme BLAST [18] dans ses calculs. Il compare les isolats testés avec ceux ayant le même spoligotype dans la base de données mondiale de marqueurs moléculaires de la tuberculose SITVIT [19], qui regroupe les données épidémiologiques [20] associées à des isolats de même spoligotype.

SpoTyping utilise des fichiers de séquences de reads simples ou par paires au format FASTQ et des fichiers de séquences complètes de génomes ou de contigs assemblés au format FASTA. Le séquences de reads sont regroupés en une unique séquence continue au format FASTA pour être ensuite soumis à l'algorithme BLAST qui détecte les régions similaires. Finalement la base de données SITVIT permet d'identifier les isolats ayant le même spoligotype. SpoTyping est limité à une lecture de 250 Mbp au sein des séquences de reads testées, lors de l'utilisation du swift mode qui accélère le temps de traitement.

SpoTyping propose un rapport statistique permettant de résumer le rapprochement avec les spoligotypes trouvés dans la base de données SITVIT, ainsi qu'une estimation du nombre de correspondances positives pour chaque spacer.

L'étude de Iwai H : et al. [11] montre l'intérêt d'une analyse de MTBC à l'aide de serveurs, appelée CASTB, et notamment le spoligotyping. Le Webserver fournit une vue complète des données, mais les performances de chaque outil utilisé ne sont pas décrites dans l'article. Il est probable que le spoligotyping prenne plus de temps en passant par un serveur suite au problème de disponibilité des données et aux lenteurs de téléchargement de ces données. Il semblerait que SpoTyping, de par sa configuration locale, puisse fournir un résultat en une minute.

D'après le repository https://github.com/xiaeryu/SpoTyping-v2.0/blob/master/SpoTyping-v2.0-commandLine/SpoTyping-README.pdf, les spécifications techniques sont les suivantes :

SpoTyping peut s'exécuter sur les principaux systèmes d'exploitation, contrairement à SpolPred qui utilise exclusivement Linux. Il se présente à la fois sous forme de script et sous forme d'application avec une interface graphique.

SpoTyping est un logiciel open-source qui peut se télécharger gratuitement à l'adresse https://github.com/xiaeryu/SpoTyping-v2.0. SpoTyping nécessite l'utilisation de Python2.7 et BLAST.

Le fichier obtenu propose une prédiction de spoligotype au format de code binaire et octal. Le fichier log obtenu contient de nombre de correspondances positives des résultats de BLAST pour chaque séquence de spacer. Le fichier xls Excel obtenu fournit le resultat de la recherche de spoligotype auprès de la base de données SITVIT WEB. (actuellement SITVIT2)

18. **BLAST Basic Local Alignment Search Tool** : logiciel basé sur l'algorithme du même nom qui détecte des régions similaires entre plusieurs séquences biologiques. Le programme compare les séquences de nucléotides aux séquences contenues dans la base de données BLAST pour fournir des résultats statistiquement significatifs.

19. La base de données SITVIT comprend les spoligotypes de tuberculosis, ainsi que les marqueurs utilisés pour les détecter MIRU12, VNTR, SIT, MIT, VIT, les différentes branches de MTBC, les pays d'origine et l'année de découverte.

20. **Epidémiologie** : discipline scientifique qui étudie les problèmes de santé dans les populations humaines, leur fréquence, leur géographie ainsi que les facteurs influants.

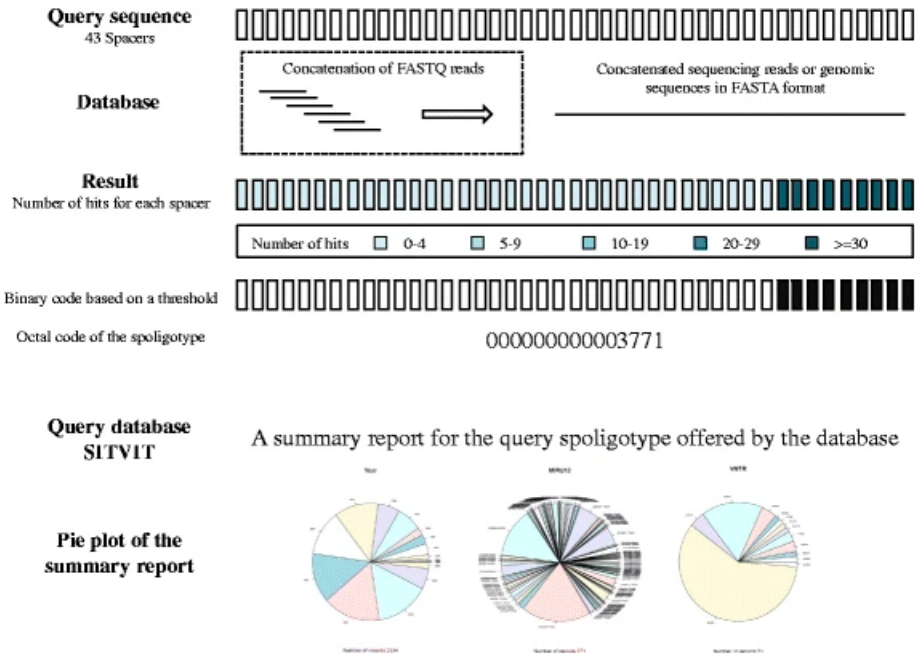
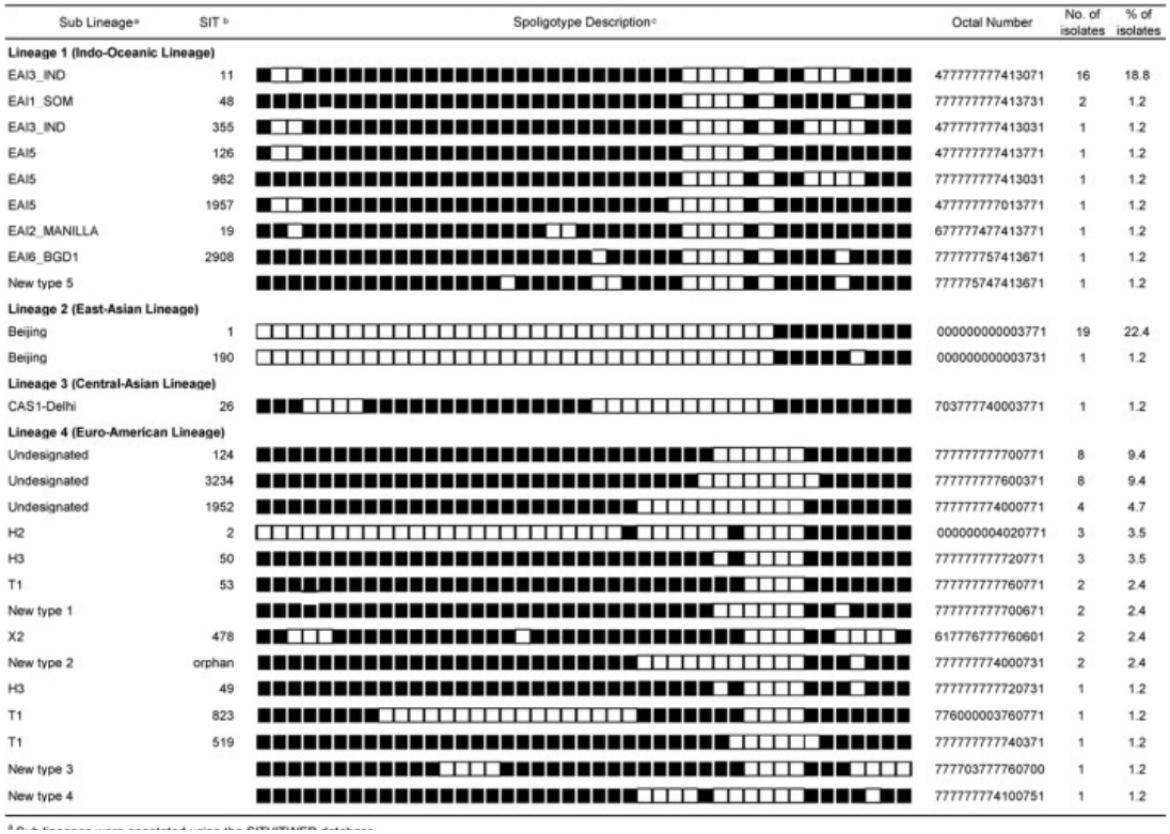


Figure 4.7 – Exemple de fonctionnement de SpoTyping, d'après *SpoTyping : fast and accurate in silico Mycobacterium spoligotyping from sequence reads*

Il est recommandé d'utiliser le swift mode paramétré par défaut si le débit de séquençage[21] est inférieur à 135 Mbp. Pour les débits de séquençage inférieurs à 135 Mbp ou supérieurs à 1260 Mbp, les seuils doivent être réglés entre 0.018 et 0.1486 fois la profondeur de lecture estimée pour les hits sans erreur, et entre 0.018 et 0.1488 fois la profondeur de lecture estimée pour les hits tolérant une erreur. Notons que la profondeur de lecture est définie par le débit de séquençage divisé par 4 500 0000 qui correspond à l'estimation de la longueur d'un génome de MTBC.

4.7. Comparaison

21. **Séquençage du génome** : consiste, par des méthodes chimiques ou de biologie moléculaire, à déterminer l'ordre des nucléotides de l'ADN.



Sub Lineage[a]	SIT [b]	Spoligotype Description[c]	Octal Number	No. of isolates	% of isolates
Lineage 1 (Indo-Oceanic Lineage)					
EAI3_IND	11		477777777413071	16	18.8
EAI1_SOM	48		777777777413731	2	1.2
EAI3_IND	355		477777777413031	1	1.2
EAI5	126		477777777413771	1	1.2
EAI5	962		777777777413031	1	1.2
EAI5	1957		477777777013771	1	1.2
EAI2_MANILLA	19		677777477413771	1	1.2
EAI6_BGD1	2908		777777757413671	1	1.2
New type 5			777775747413671	1	1.2
Lineage 2 (East-Asian Lineage)					
Beijing	1		000000000003771	19	22.4
Beijing	190		000000000003731	1	1.2
Lineage 3 (Central-Asian Lineage)					
CAS1-Delhi	26		703777740003771	1	1.2
Lineage 4 (Euro-American Lineage)					
Undesignated	124		777777777700771	8	9.4
Undesignated	3234		777777777600371	8	9.4
Undesignated	1952		777777774000771	4	4.7
H2	2		000000004020771	3	3.5
H3	50		777777777720771	3	3.5
T1	53		777777777760771	2	2.4
New type 1			777777777700671	2	2.4
X2	478		617776777760601	2	2.4
New type 2	orphan		777777774000731	2	2.4
H3	49		777777777720731	1	1.2
T1	823		776000003760771	1	1.2
T1	519		777777777740371	1	1.2
New type 3			777703777760700	1	1.2
New type 4			777777774100751	1	1.2

Figure 4.8 – Exemple comparatif de 26 spoligotypes correspondant à 4 lignées différentes de M. tuberculosis, d'après *Insight into genetic diversity of Mycobacterium tuberculosis in Kandy, Sri Lanka reveals predominance of the Euro-American lineage*

Références

[1] Comas I. et al. *Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans*

[2] Coll F. et al., *SpolPred : rapid and accurate prediction of Mycobacterium tuberculosis spoligotypes from short genomic sequences. Bioinformatics. 2012 ;28(22) :2991–3*

[3] Brynildsrud O.B. et al., *Global expansion of Mycobacterium tuberculosis lineage 4 shaped by colonial migration and local adaptation*

[4] Driscoll J. R., *Spoligotyping for molecular epidemiology of the Mycobacterium tuberculosis complex*

[5] Jinek M. et al, *A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity*

[6] Gori A. et al, *Spoligotyping and Mycobacterium tuberculosis*

[7] Perry S. et al., *Infection with Helicobacter pylori is associated with protection against tuberculosis*

[8] Perry S. et al, *The immune response to tuberculosis infection in the setting of Helicobacter pylori and helminth infections*

[9] Xia E. et al., *SpoTyping : fast and accurate in silico Mycobacterium spoligotyping from sequence reads*

[10] Dale JW. et al., *Spacer oligonucleotide typing of bacteria of the Mycobacterium tuberculosis complex : recommendations for standardised nomenclature. Int J Tuberc Lung Dis. 2001 ;5(3) :216–9*

[11] Iwai H et al., *CASTB (the comprehensive analysis server for the Mycobacterium tuberculosis complex) : A publicly accessible web server for epidemiological analyses, drug-resistance prediction and phylogenetic comparison of clinical isolates. Tuberculosis. 2015 ;95(6) :843–4*

[12] Demay C. et al., *SITVITWEB - A publicly available international multimarker database for studying Mycobacterium tuberculosis genetic diversity and molecular epidemiology. Infect Genet Evol. 2012 ;12 :755–66*

[13] McGovern Institute Channel, *Genome Editing with CRISPR-Cas9*

[14] O'Neil M.B. et al., *Lineage specific histories of Mycobacterium tuberculosis dispersal in Africa and Eurasia* bioRxiv 10.1101/210161

[15] Ratovonirina N. H. *Etudes descriptive, épidémiologique, moléculaire et spatiale des souches Mycobacterium tuberculosis circulant à Antananarivo, Madagascar*

[16] Kamerbeek et al., *Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology*

[17] , *Insight into genetic diversity of Mycobacterium tuberculosis in Kandy, Sri Lanka reveals predominance of the Euro-American lineage*