Exploiter l'intégralité des données de séquençage de Mycobacterium tuberculosis : une plateforme pour l'analyse génomique in-silico à grande échelle

Gaëtan Senelle

THESE DE DOCTORAT DE L'ETABLISSEMENT UNIVERSITE BOURGOGNE FRANCHE-COMTE

PREPAREE A L'UNIVERSITE DE FRANCHE-COMTE

Ecole doctorale n°37

Ecole Doctorale de Sciences Physiques pour l'Ingénieur et Microtechniques

Doctorat d'informatique

Par

M. SENELLE Gaëtan

Exploiter l'intégralité des données de séquençage de *Mycobacterium tuberculosis*: une plateforme pour l'analyse génomique in-silico à grande échelle

Thèse présentée et soutenue à Belfort, le 18 décembre 2024

Composition du Jury :

M. Francois-Xavier WEILL	Professeur, Institut Pasteur	Président, Rapporteur
M. Stéphane CHRÉTIEN	Professeur, Université Lyon 2	Rapporteur
M. David LAIYMANI	Maître de conférences, Université de Franche-Comté	Examinateur
Mme. Claire TOFFANO-NIOCHE	Chercheur CNRS, Université Paris-Saclay	Examinatrice
M. Christophe GUYEUX	Professeur, Université de Franche-Comté	Directeur de thèse
M. Christophe SOLA	Professeur, Université Paris-Saclay	Codirecteur de thèse

Remerciements

Je tenais tout d'abord à remercier mon directeur de thèse, le professeur Christophe GUYEUX, pour son accompagnement tout au long de cette thèse. Son enthousiasme pour ce projet m'a poussé à aller toujours plus loin dans nos travaux. Je resterai marqué par son dévouement, son investissement et sa présence tout au long de nos recherches. Sa participation a donné une autre dimension à notre travail et a permis des découvertes que, sans lui, nous n'aurions pu réaliser.

Je souhaitais également remercier mon co-directeur de thèse, le professeur Christophe SOLA dont la présence, l'expertise et la passion pour le sujet m'ont été d'une aide indispensable pour mener ce travail. Merci pour tout ce que vous m'avez permis d'apprendre sur la biologie, la tuberculose et les mycobactéries.

Merci aussi au Dr. Guislaine REFRÉGIER pour son accompagnement scientifique et son soutien durant ce travail, et notamment pour ses précieux conseils et relectures lors de la rédaction de cette thèse.

Je remercie l'ensemble des membres du jury de s'être rendu disponible pour la soutenance de cette thèse. Merci au professeur Francois-Xavier WEILL et au professeur Stéphane CHRÉTIEN d'avoir accepté d'être rapporteurs, et au Dr. David LAIYMANI et au Dr. Claire TOFFANO-NIOCHE d'avoir accepté d'être examinateurs.

Je voulais aussi remercier Dr. Rodolphe BOUDOT et Pr. Didier HOCQUET pour leurs conseils et pour avoir accepté de prendre part au comité de suivi individuel de cette thèse.

Enfin, je souhaitais remercier mes parents ainsi que ma compagne, Amélia, pour m'avoir toujours soutenu, conseillé et encouragé tout au long de ce doctorat.

Acronymes

ADN acide désoxyribonucléique. 15, 16, 18–20, 22–24, 26–28, 30–32, 49, 51–55, 64

API Application Programming Interface. 72, 77

ARN acide ribonucléique. 15–17, 19, 32

ARNm acide ribonucléique messager. 17, 18

ARNr acide ribonucléique ribosomique. 17, 43

ARNt acide ribonucléique de transfert. 17

BAM binary alignment map. 55, 63

BCG Bacille Calmette-Guérin. 36, 37, 44

BP Before present. 82

BWA Burrows-Wheeler Alignment. 55

CAS CRISPR-associated. 45

CIGAR Compact Idiosyncratic Gapped Alignment Report. 53, 64

CRAM Compressed Reference-oriented Alignment Map. 55, 63, 65

CRISPR clustered regularly interspaced short palindromic repeats. 44, 45, 69, 83, 99–102

DR direct repeat. 45, 69

DVR-PCR direct variable repeat polymer chain reaction. 45

ETR exact tandem repeat. 48, 49

GPT Générative Pre-trained Transformer. 102, 103

HNSW hierarchical navigable small world. 104

IR répétition inversée. 23

IS séquence d'insertion. 23, 24, 43, 63, 68, 69, 71–73, 75, 99, 140

LLM large language model. 102, 103, 106

MDR-TB tuberculose multirésistante. 38

MIRU mycobacterial interspersed repetitive unit. 49, 50

MPTR major polymorphic tandem repeat. 48

MTBC *Mycobacterium tuberculosis* complex. 12, 13, 32, 39, 42–45, 48–51, 54, 55, 59–63, 65, 68, 70, 75, 77–82, 84, 89, 92, 95, 99–102, 107, 139

MVR-PCR minisatellite variant repeat polymer chain reaction. 45

NCBI National Center for Biotechnology Information. 51, 96, 102

NGS Séquençage de nouvelle génération. 24, 28, 51, 52, 59, 60, 78, 79

PCR réaction en chaîne par polymérase. 26–28, 43, 45, 46, 48, 63, 65

qRT-PCR réaction en chaîne par polymérase en temps réel. 27

RAG Retrieval Augmented Generation. 103
RD region of difference. 43, 44, 65–67, 71–73, 82, 92, 95, 99
RFLP polymorphisme de longueur des fragments de restriction. 43, 49
RR-TB tuberculose résistante à la rifampicine. 38

SAM sequence alignment map. 55
SNP single-nucleotide polymorphism. 20, 39, 49, 50, 56, 57, 63, 64, 69, 72–75, 78, 81, 83–90, 92–97, 99–101, 140, 141
SNV single-nucleotide variant. 20
STB smooth tuberculosis bacilli. 43

tf-idf term frequency–inverse document frequency. 74

UEP unique event polymorphism. 44
UMI unique molecular identifiers. 62

VCF Variant Call Format. 63
VNTR variable number tandem repeat. 48, 49

WGS whole-genome sequencing. 8, 49–51, 60, 78, 79, 82, 83, 95, 96

Table des matières

Chapitre 1

Introduction

La tuberculose demeure un problème de santé publique majeur à l'échelle mondiale. Selon l'Organisation mondiale de la santé (OMS), en 2022, la tuberculose était la deuxième cause de décès dans le monde due à un seul agent infectieux, après le COVID-19, et causait presque deux fois plus de décès que le VIH. Plus de 10 millions de personnes continuent de contracter la tuberculose chaque année. En 2022, le nombre total de décès causés par la tuberculose (y compris ceux chez les personnes vivant avec le VIH) dans le monde était de 1,30 million [205].

De plus, l'émergence de formes de tuberculose résistantes aux médicaments est une préoccupation croissante, car certaines souches sont beaucoup plus difficiles à traiter et peuvent entraîner des taux de mortalité plus élevés. Étudier la diversité mondiale de cet agent infectieux est crucial pour comprendre la maladie, développer de nouveaux traitements, et tenter de contrôler et éliminer la menace que constitue la tuberculose pour la santé publique.

Depuis plus d'une décennie, l'étude de cette maladie peut également se faire par l'analyse des génomes obtenus de nombreuses souches de tuberculose. Disposer de nombreux génomes diversifiés est intéressant pour plusieurs raisons. D'une part, cela permet une compréhension de la diversité génétique de la tuberculose. Il existe plusieurs sous-espèces avec des adaptations écologiques variées que l'on retrouve sous forme d'information génomique. Il est ainsi possible d'identifier les variations génétiques spécifiques qui existent au sein des populations. Cela fournit des informations sur l'évolution et la propagation de la tuberculose, ainsi que sur son adaptation à différents environnements hôtes. D'autre part, l'étude du génome de la tuberculose aide à la compréhension de l'émergence et de la propagation des souches de tuberculose résistantes aux médicaments, qui sont aujourd'hui un problème majeur dans la lutte contre la maladie. Il est possible, pas comparaison entre de multiples génomes, de déterminer les caractéristiques génomiques responsables de la sensibilité ou de la résistance aux médicaments. Cela permet l'étude de nouvelles stratégies de traitement à grande échelle mais aussi la personnalisation des traitements pour une infection donnée. La compréhension du génome aide aussi à identifier les gènes et protéines qui peuvent être ciblés par des vaccins et ainsi augmenter les chances de développer un vaccin efficace sur des souches diverses de tuberculose. Une meilleure compréhension génomique permet aussi d'aider au développement de diagnostics plus précis et permettre ainsi une identification plus précise des souches.

Problèmatique

Avec le développement des méthodes de séquençage et la réduction de leur coût, on compte actuellement plus de 160 000 séquences de génomes disponibles publiquement et collectées sur la plateforme du National Center for Biotechnology Information. Il s'agit d'une source de données d'une grande richesse mais qui reste une source de données brute, qu'il reste à analyser. Il s'agit là d'un défi majeur reconnu, de par l'absence de standardisation des pipelines d'analyse et le manque de bases de données pour partager et étudier

les données analysées à un niveau mondial [180]. Exploiter toute la donnée génomique brute requiert l'intégration de connaissances et de techniques très variées. D'une part, il est nécessaire de regrouper la connaissance génomique issue d'années d'études dans la littérature scientifique, et donc d'avoir la participation de scientifiques experts du domaine. D'autre part, il est nécessaire ensuite de développer des algorithmes et pipelines bioinformatiques pour extraire les marqueurs génomiques d'intérêt parmi la quantité de données brutes. Si des outils existent pour étudier des caractéristiques spécifiques des génomes, ils montrent leurs premières limites quand il s'agit d'analyser l'intégralité des caractéristiques génomiques. Cela s'explique pas la diversité des méthodes utilisées et le manque de standardisation autour de ces méthodes. Enfin, quand il s'agit de l'analyse de l'intégralité des génomes disponibles, les contraintes techniques se multiplient. L'assemblage d'outils spécifiques n'est pas compatible avec la charge de données, il est nécessaire de repenser les outils dans leur globalité. La masse de données impose aussi de concevoir des outils d'analyse adaptés, avec la mise en place d'infrastructures de serveurs, de pipeline d'ingestion de données, de base de données avec des capacités d'analyses avancées sur une masse d'informations très importante.

La prise en compte des contraintes techniques évoquées précédemment est essentielle, mais bien souvent la question de l'accessibilité des outils est ignorée. De tels outils peuvent être utilisés aussi bien par des bioinformaticiens que par des biologistes, et ce pour une grande variété d'études qui va de la phylogénie à la résistance aux médicaments. Concevoir des outils complexes à exploiter et nécessitant des compétences techniques avancées implique de se priver de toute une partie des acteurs de la recherche sur la tuberculose. C'est pour cela que les nouveaux outils se doivent d'être facilement accessibles sans nécessiter de matériel spécifique ou de connaissances bioinformatiques avancées.

Objectifs

Ce travail propose une approche globale et à grande échelle pour l'étude et l'exploitation de la donnée génomique brute disponible sur la tuberculose, et ce dans son ensemble et sa finesse. Il s'agit de proposer une résolution dans l'étude de marqueurs génétiques qui soit similaire à des outils spécifiques, mais en plus de réaliser ce travail sur l'intégralité de la donnée disponible.

Cette approche globale nécessite plusieurs étapes. Dans un premier temps, il s'agit de répertorier toute la connaissance sur les caractéristiques du génome de la tuberculose que l'on peut trouver dans des décennies de littérature scientifique. Ensuite, à partir de ces éléments, de développer de nouveaux outils d'analyse génomique *in-silico* compatibles avec une approche globale. Enfin, de développer et rendre accessible des systèmes d'analyse de ces données aussi bien statistiques que phylogéniques.

Présentation du plan

Ce travail étant situé au croisement de l'informatique et de la biologie, une première partie introduit les pré-requis en termes de biologie et de génomique. Une attention particulière a été apportée à ces éléments dans le but de synthétiser la connaissance acquise sur le sujet durant cette thèse, mais aussi de fournir une introduction complète pour la compréhension de la suite du travail. Il est abordé les bases de la génomique, et il sera fait une description des principales variations génétiques et les différentes méthodes d'analyses du génome. Un niveau de détail important est parfois donné car il permet de comprendre certaines caractéristiques génomiques que l'on peut retrouver plus tard au sein des analyses bioinformatiques. On introduit ensuite la turberculose, son histoire et l'état de la lutte contre la maladie dans le monde. La troisième moitié de cette partie est consacrée au *Mycobacterium tuberculosis* complex (MTBC) qui est le groupe de mycobactéries causant la tuberculose. Nous présenterons en détails les caractéristiques génomiques du MTBC connues. Enfin nous présenterons les outils et méthodes de bioinformatique couramment

utilisés pour l'étude du génome de la tuberculose.

La deuxième partie est consacrée à la présentation de la plateforme d'analyse TB-Annotator, principale contribution de cette thèse, qui se divise en deux travaux principaux. D'une part, le pipeline d'analyse, qui est l'ensemble de programmes et d'outils permettant la transformation des données génomiques brutes en caractéristiques génomiques spécifiques aux génomes de la tuberculose. Nous présenterons toutes les nouvelles méthodes et nouveaux algorithmes développés à cette occasion. Nous verrons en quoi ce pipeline est suffisamment performant pour permettre l'analyse de données massives et très diversifiées. D'autre part, nous présenterons la plate-forme d'analyse web complémentaire au pipeline. Il sera abordé les capacités de calculs de la plateforme sur l'ensemble des marqueurs génomiques collectés par le pipeline, ainsi que la capacité d'analyse phylogénétique. Enfin nous terminerons avec une comparaison de divers outils existants.

La troisième partie présente trois études qui correspondent à trois apports pour la taxonomie du MTBC, en se basant sur le travail réalisé avec TB-Annotator. La première étude relate la découverte d'une sous-lignée de la lignée 2 du MTBC. La seconde étude propose différents nouveaux marqueurs phylogénétiques pour la lignée 4, la lignée 6 et la lignée 7. Enfin, la troisième étude relate une découverte importante pour la phylogénie du MTBC avec l'identification d'une nouvelle lignée rare, la lignée 10. Dans ces trois études, TB-Annotator a permis de dépasser l'état de la recherche de par son approche globale et l'échelle des analyses effectuées.

La quatrième partie conclut ce travail en proposant une vision du futur de TB-Annotator, sur deux axes. D'une part, sur les possibilités d'amélioration et d'extension de la plateforme. Et d'autre part, sur les possibilités d'exploitation de la masse de données encore sous exploitée que constitue l'ensemble de la littérature scientifique sur le sujet. Incorporer une part plus importante de cette littérature dans la plateforme pourrait ouvrir la voie à de nombreuses autres avancées, notamment dans la compréhension de la résistance aux médicaments.

Liste des contributions

1. *Tb-annotator : a scalable web application that allows in-depth analysis of very large sets of publicly available Mycobacterium tuberculosis complex genomes*. Senelle, Guyeux, Refrégier, and Sola [240] (2023).
2. *Connection between two historical tuberculosis outbreak sites in japan, honshu, by a new ancestral mycobacterium tuberculosis l2 sublineage*. Guyeux, Senelle, Refrégier, Bretelle-Establet, Cambau, and Sola [108] (2022).
3. *Towards a better understanding of the long-lasting evolutionary history of mycobacterium tuberculosis*. Senelle, Guyeux, Refrégier, and Sola [241] (2023).
4. *Newly identified mycobacterium africanum lineage 10, central africa*. Guyeux, Senelle, Meur, Supply, Gaudin, Phelan, Clark, Rigouts, de Jong, Sola, and Refrégier [110] (2024).

Chapitres d'ouvrages et autres contributions :

5. *Investigating the diversity of tuberculosis spoligotypes with dimensionality reduction and graph theory*. Senelle, Guyeux, Refrégier, and Sola [237] (2022).
6. *An updated evolutionary history and taxonomy of Mycobacterium tuberculosis lineage 5, also called M. africanum*. Sahal, Senelle, La, Molina-Moya, Dominguez, Panda, Cambau, Refregier, Sola, and Guyeux [230] (2022).
7. *Mycobacterium tuberculosis complex drug-resistance, phylogenetics, and evolution in Nigeria : Comparison with Ghana and Cameroon*. Sahal, Senelle, La, Panda, Taura, Guyeux, Cambau, and Sola [231] (2023).
8. *Evolution, Phylogenetics, and Phylogeography of Mycobacterium tuberculosis complex*. Sola, Mokrousov, Sahal, La, Senelle, Guyeux, Refrégier, and Cambau [255] (2024).
9. *Tools for short variant calling and the way to deal with big datasets*. Meur, Zein-Eddine, Lamer, Hak, Senelle, Vernadet, O'Donnell, de la Vega, and Refrégier [183] (2024).

Articles de conférences :

10. *Le chaos croissant de la génomique des populations des bacilles de la tuberculose à l'ère des big data*. Sola, Guyeux, and Senelle [253] (2022).
11. *The growing chaos of tuberculosis population genomics at the era of «big data» : sorting out the wheat from the chaff*. Sola, Senelle, La, Billard-Pomares, Marin, Bridier-Nahmias, Guyeux, Refrégier, Carbonnelle, and Cambau [254] (2022).
12. *Tb-annotator : A scalable web application that allows in-depth analysis of very large sets of publicly available Mycobacterium tuberculosis complex genomes*. Senelle, Guyeux, Cambau, Refrégier, and Sola [239] (2023).
13. *The hidden diversity of Mycobacterium tuberculosis complex in africa : the new l10 and the possible diversification histories of the complex*. Guyeux, Senelle, Le Meur, Sola, and Refrégier [109] (2024).

Chapitre 2

Éléments de génomique et de bio-informatique

2.1 Génomique

2.1.1 Le génome

Le génome est constitué de l'ensemble des informations génétiques d'un organisme. Ces informations sont codées par de longues séquences d'acide désoxyribonucléique (ADN), à l'exception de certains virus dont le génome est constitué d'acide ribonucléique (ARN).

Le génome est majoritairement formé de chromosomes, dont le nombre varie en fonction de l'espèce. Chaque chromosome est composé d'une unique molécule d'ADN, qui est de forme linéaire chez les eucaryotes et habituellement circulaire chez les procaryotes. Chez la plupart des bactéries, et notamment chez *Mycobacterium tuberculosis*, le chromosome est unique et circulaire.

Mis à part l'ADN chromosomique, le génome est aussi formé de l'ADN des plasmides et de l'ADN présent dans les organites chez les eucaryotes (comme les mitochondries et les chloroplastes)[152].

Les premières preuves que le génome est fait d'ADN furent obtenues entre 1945 et 1952, mais c'est en 1953, avec la découverte de la structure en double-hélice par Watson et Crick, que les biologistes furent convaincus que l'ADN était bien le matériel génétique [27].

2.1.2 L'ADN

L'acide désoxyribonucléique est un polymère, linéaire et sans branche, où chaque monomère est appelé nucléotide. Chaque molécule d'ADN est constituée de deux polynucléotides liés ensemble par paires de nucléotides (appelées paires de bases) avec des liaisons hydrogène, et formant ainsi une double hélice (figure 2.1). La longueur de cette chaîne peut aller de quelques centaines de paires à plusieurs millions de paires de bases. Le diamètre de la double hélice est de 20 ångström. Les liaisons hydrogène sont dites faibles, ce qui peut permettre la séparation des brins.

Chaque nucléotide est constitué de 3 composants :

— Une molécule de désoxyribose qui est un pentose, un sucre composé de 5 atomes de carbone. Les atomes de carbone sont numérotés de $1'$ à $5'$ (à lire «1 prime à 5 prime»).

— Une base azotée (ou base nucléique) qui peut être cytosine, adénine, thymine ou guanine. Cette base est attachée au carbone $1'$ du désoxyribose.

— Un groupe phosphate avec entre une et trois molécules de phosphate (monophosphate, diphosphate, triphosphate). Le groupe est attaché au carbone $5'$ du désoxyribose.

Les nucléotides individuels sont reliés entre eux par des liaisons phosphodiester entre leurs carbones $5'$ et $3'$. Le polynucléotide a donc une direction chimique, exprimée $5' \rightarrow$

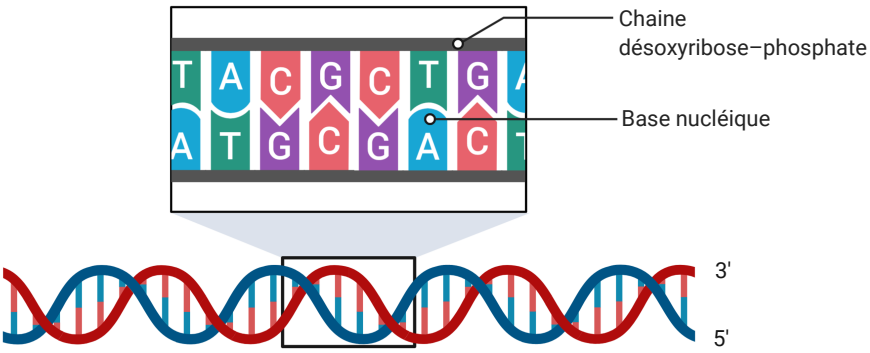


FIGURE 2.1 – Structure de l'ADN

3′ sur une des chaînes, et 3′ → 5′ sur la chaîne opposée. L'ADN polymérase, qui est une enzyme qui synthétise l'ADN, le fait toujours dans la direction 5′ → 3′ [27]. Par convention on écrit les séquences d'ADN dans le sens 5′ → 3′.

L'adénine et la guanine sont des purines, la cytosine et la thymine sont des pyrimidines. La thymine forme une paire de bases avec l'adénine, et la cytosine forme une paire de bases avec la guanine (figure 2.2). Cela a pour conséquence que si la séquence des bases est connue sur l'une des chaînes, alors la séquence sur l'autre peut être déterminée [296].

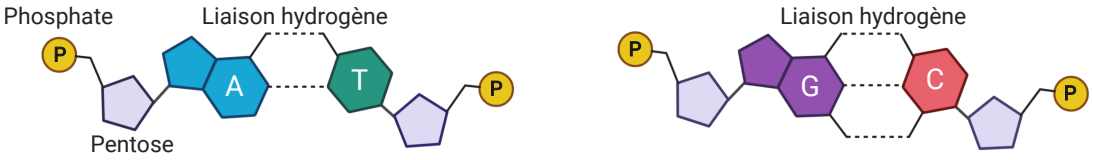


FIGURE 2.2 – Les deux paires de bases de l'ADN

2.1.3 Expression du génome

Le génome ne participe pas activement au développement d'un organisme, il est seulement porteur de l'information génétique et n'est pas capable de le transmettre aux cellules. Par une série d'interactions complexes, appelées expression génétique, ressortent deux produits du génome que sont le transcriptome et le protéome [27].

Transcriptome

Le transcriptome est l'ensemble des molécules d'ARN dérivé du génome, et maintenu par un processus appelé transcription. Des enzymes appelées ARN polymérase sont responsables de la transcription des séquences d'ADN, appelées gènes, en ARN [27, 152].

Structurellement, l'ARN est un polynucléotide similaire à l'ADN, à la différence que le sucre est le ribose et que l'ARN possède, à la place de la thymine, l'uracile. Les molécules d'ARN font rarement plus de quelques milliers de nucléotides de longueur et la majorité est sous la forme d'un simple brin, contrairement à l'ADN.

La transcription s'effectue à partir d'ADN servant de modèle pour la séquence d'ARN synthétisée. L'ADN est lu dans la direction 3′ → 5′, et l'ARN est synthétisé dans le sens 5′ → 3′ avec les bases nucléiques complémentaires à celles de l'ADN lu.

L'ARN contenu dans une cellule peut occuper différentes fonctions. On distingue d'une part l'ARN codant de l'ARN non-codant.

Une seule classe de molécules constitue l'ARN codant, l'acide ribonucléique messager (ARNm). La transcription des gènes dont ils sont issus sert d'intermédiaire dans la seconde phase de l'expression du génome, où ARNm sera traduit en protéines constituantes du protéome. L'ARNm constitue rarement plus de 4% de l'ARN total et a une durée de vie très courte. Par exemple, celui des bactéries a une demi-vie de seulement quelques minutes.

L'ARN non-codant, lui, n'est pas traduit en protéines. On en distingue plusieurs types, dont deux particulièrement importants.

— L'acide ribonucléique ribosomique (ARNr) qui est le plus abondant et qui constitue plus de 80% de l'ARN total. Ce sont des composants des ribosomes où la synthèse des protéines a lieu.
— L'acide ribonucléique de transfert (ARNt) participe aussi à la synthèse des protéines. Il a pour rôle d'approvisionner les ribosomes en acides aminés et de s'assurer que ces acides aminés sont liés dans l'ordre spécifié par la séquence d'ARNm qui est traduite [27].

Protéome

Le protéome est l'ensemble des protéines de la cellule qui sont synthétisées par un processus appelé traduction. Ces protéines définissent les réactions biochimiques possibles au sein de la cellule, ou peuvent aussi avoir un rôle structurel. Les protéines sont synthétisées par la traduction des molécules d'ARNm du transcriptome en polypeptides.

Les polypeptides sont des polymères constitués d'une chaîne linéaire, sans branche, d'acides aminés liés par des liaisons dites «peptidiques», qui sont des liaisons covalentes. Leur longueur dépasse rarement les 2000 acides aminés.

On peut décrire la structure d'une protéine selon 4 niveaux hiérarchiques :

— **La structure primaire** de la protéine qui est constituée par la chaîne d'acides aminés formant le polypeptide.
— **La structure secondaire** qui correspond aux différentes dispositions dans l'espace que peut adopter le polypeptide. Les deux configurations les plus communes sont les hélices α et les feuillets β. Pour les hélices α, la chaîne polypeptidique est de forme hélicoïdale, et pour les feuillets β la chaîne forme un plan plissé (similaire à un accordéon). Ces structures sont stabilisées principalement du fait des liaisons hydrogène qui se forment entre les acides aminés de la chaîne.
— **La structure tertiaire** correspond au repliement dans l'espace en 3 dimensions de la chaîne polypeptidique. Cette structure est stabilisée par de multiples forces chimiques (liaisons hydrogène, liaisons ioniques, interactions hydrophobes, liaisons de Van der Waals et parfois des liaisons covalentes)
— **La structure quaternaire** est l'association de plusieurs polypeptides, ou sous-unités, qui forment une protéine multimérique (qui signifie qu'elle possède plusieurs sous-unités). Ces structures quaternaires peuvent être stables grâce à des liaisons fortes (pont disulfure), alors que d'autres moins bien stabilisées (interactions hydrophobes, liaisons hydrogène) peuvent avoir des changements de sous-unités ou revenir à leur forme polypeptidique.

Au final, des séquences différentes d'acides aminés entraînent des combinaisons variées de réactivités chimiques et ainsi une diversité fonctionnelle des protéines. Ces combinaisons ne déterminent pas seulement la structure globale de la protéine résultante, mais également le positionnement des groupes réactifs à la surface de la structure, lesquels établissent les propriétés chimiques de la protéine.

Code génétique

C'est le code génétique qui spécifie la correspondance entre la séquence d'ARNm et la séquence d'acides aminés qui constitue la protéine. Un triplet de nucléotides d'ARNm,

Acide aminé	Code 3 lettres	Code 1 lettre	Codons
Alanine	Ala	A	GCU, GCC, GCA, GCG
Arginine	Arg	R	CGU, CGC, CGA, CGG ; AGA, AGG
Asparagine	Asn	N	AAU, AAC
Acide aspartique	Asp	D	GAU, GAC
Cystéine	Cys	C	UGU, UGC
Glutamine	Gln	Q	CAA, CAG
Acide glutamique	Glu	E	GAA, GAG
Glycine	Gly	G	GGU, GGC, GGA, GGG
Histidine	His	H	CAU, CAC
Isoleucine	Ile	I	AUU, AUC, AUA
Leucine	Leu	L	UUA, UUG ; CUU, CUC, CUA, CUG
Lysine	Lys	K	AAA, AAG
Méthionine	Met	M	AUG
Phénylalanine	Phe	F	UUU, UUC
Proline	Pro	P	CCU, CCC, CCA, CCG
Serine	Ser	S	UCU, UCC, UCA, UCG ; AGU, AGC
Threonine	Thr	T	ACU, ACC, ACA, ACG
Tryptophan	Trp	W	UGG
Tyrosine	Tyr	Y	UAU, UAC
Valine	Val	V	GUU, GUC, GUA, GUG

TABLE 2.1 – Liste des 20 acides aminés et leurs codons dans le code génétique

appelé codon, est associé à un acide aminé parmi les 20 différents existants (tableau 2.1). Plusieurs codons peuvent désigner le même acide aminé, on parle alors de codons synonymes. Le code a des codons spécifiques pour signifier le point de départ dans l'ARNm où la traduction commence et se termine. Le codon d'initiation est en général AUG qui correspond aussi à la méthionine, d'autres codons peuvent être utilisés comme UUG notamment dans les bactéries. Les codons de terminaison sont eux UAG, UAA et UGA. Il est important de noter que le code génétique n'est pas universel et qu'il existe des exceptions.

Lors de la traduction, les nucléotides sont lus 3 par 3, sans chevauchement. Pour toute séquence de nucléotides, il existe 3 manières différentes de lire la séquence en fonction du point de départ, on parle de différents cadres de lecture. Chaque séquence a donc 3 cadres de lecture possibles avec 3 séquences de codons qui spécifient des séquences d'acides aminés différentes. On parle de cadre de lecture ouvert pour le cadre de lecture commençant par un codon d'initiation, se terminant par un codon de terminaison et encodant une séquence d'acides aminés relativement longue (> 50 codons). On parle de cadre de lecture fermé quand la séquence ne peut être traduite en polypeptide.

Bien que les séquences d'ARNm soient traduites en protéines, elles ne représentent pas à elles seules la diversité des protéines. Par exemple, deux acides aminés rares, la sélénocystéine et la pyrrolysine, peuvent parfois être insérés lors de la synthèse des protéines. Les protéines sont donc spécifiées par le génome, mais aussi par des variations qui peuvent se produire lors de la synthèse [27].

2.2 Variation génétique

2.2.1 Réplication de l'ADN

Afin de maintenir l'information génétique d'un organisme, il est crucial que l'ADN puisse se répliquer, et ce de manière exacte. Lors de la réplication, une molécule d'ADN parent se transforme en deux molécules d'ADN filles. Chaque molécule d'ADN enfant est constituée d'un polynucléotide provenant du parent et d'un polynucléotide complémentaire

nouvellement synthétisé. Ce type de réplication est appelé réplication semi-conservative.

La réplication débute à un point de l'ADN appelé «Origine de réplication». Ce point est unique chez les bactéries mais de multiples points peuvent exister chez les eucaryotes. La protéine DnaA reconnait l'origine de réplication et promeut la séparation (appelée dénaturation) des deux brins d'ADN. Une hélicase sépare les deux brins en rompant les liaisons hydrogène, cela crée ainsi deux fourches de réplication. Cette séparation est temporaire car, en parallèle, une renaturation est opérée par l'ADN polymérase au niveau des deux fourches de réplications. L'ADN polymérase procède à la synthèse de l'ADN sur les deux brins et dans des directions opposées (figure 2.3).

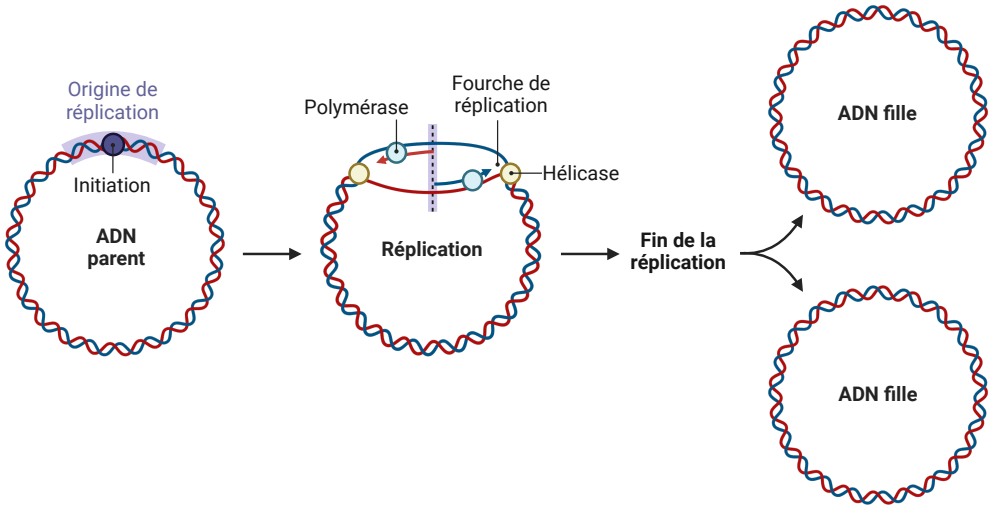


FIGURE 2.3 – Réplication de l'ADN

L'ADN polymérase possède des limitations qui complexifient la réplication. Une des principales limitations est causée par le fait que la synthèse s'effectue toujours dans le sens $5' \rightarrow 3'$. Cela a pour conséquence que, pour un des deux brins, l'ADN polymérase synthétise l'ADN de manière continue, ce brin étant appelé «brin direct». Pour l'autre brin, la synthèse doit s'effectuer par fragments (appelés fragments d'Okazaki) dans le sens $5' \rightarrow 3'$, ce brin est appelé «brin indirect».

Une autre limitation dans le processus de réplication de l'ADN est liée au fait que l'ADN polymérase ne peut pas entamer la réplication en se basant uniquement sur un brin d'ADN comme modèle. Pour démarrer la synthèse, cette enzyme nécessite une «amorce». La primase est l'enzyme responsable de la synthèse de courts segments d'ARN (10 à 12 nucléotides de longueur) qui servent d'amorces pour l'ADN polymérase (figure 2.4). Cet ARN amorce sera ensuite supprimé par une enzyme une fois la synthèse débutée.

Le processus d'amorçage ne doit se produire qu'une seule fois sur le brin direct, au sein de l'origine de réplication, car une fois amorcée, la copie du brin direct est synthétisée de manière continue jusqu'à l'achèvement de la réplication. En revanche, sur le brin indirect, l'amorçage est un processus répété qui doit avoir lieu chaque fois qu'un nouveau fragment d'Okazaki est initié.

2.2.2 Mutations

Bien que la réplication de l'ADN soit un processus particulièrement fiable, il peut arriver que des erreurs de réplication se produisent, et altèrent ainsi l'ADN. Des modifications de l'ADN peuvent aussi être causées par des mutagènes, qui sont des agents qui changent le génome (comme les radiations ou certains composés chimiques). On appelle «mutation» les changements de séquence de nucléotides d'une petite région de la molécule d'ADN. On parle de mutations spontanées lorsqu'elles apparaissent dans des conditions normales (par exemple lors de la réplication), et de mutations induites lorsqu'elles sont causées par des

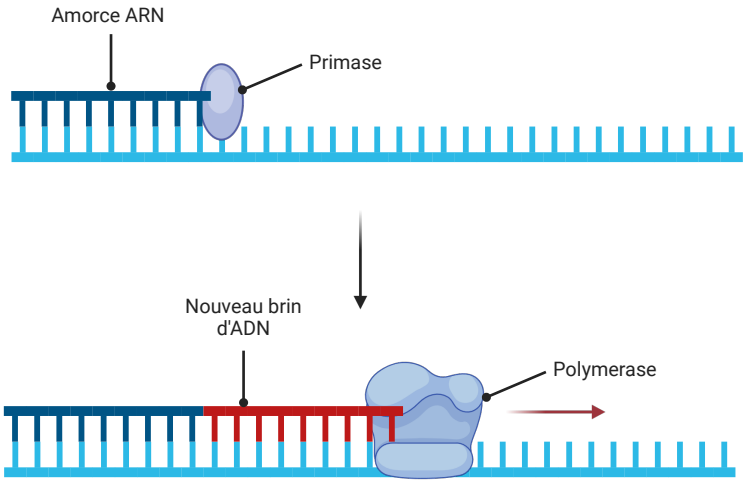


FIGURE 2.4 – Amorce lors de la synthèse de l'ADN

facteurs extérieurs, comme par des mutagènes.

SNV et SNP

Le single-nucleotide variant (SNV) est la mutation la plus simple, il s'agit du changement d'un unique nucléotide dans la séquence d'ADN. Les single-nucleotide polymorphisms (SNPs) (à lire «snip») sont un sous-type des SNV. Pour être considéré comme SNP, un SNV doit être présent dans au moins 1% de la population, bien qu'il n'y ait pas de consensus autour de ce chiffre. On utilise souvent ces termes de manière interchangeable en pratique. On peut classer ces mutations en deux catégories :

— **Les transitions** qui sont de purine vers purine, ou de pyrimidine vers pyrimidine (A → G, G → A, C → T, ou T → C).
— **Les transversions** qui sont de purine vers pyrimidine, ou de pyrimidine vers purine (A → C, A → T, G → C, G → T, C → A, C → G, T → A, or T → G).

Certaines de ces mutations sont spontanées, lors de la réplication de l'ADN par exemple. Il en existe plusieurs causes, mais nous pouvons illustrer le phénomène avec la rare tautomérisation des bases nucléiques. Par exemple, la thymine existe en deux tautomères, qui sont des variations dans l'enchaînement des atomes de thymine. Un de ces tautomères, sous forme d'énol, peut former une paire avec la guanine au lieu de l'adénine. Lors des réplications suivantes, ce rare tautomère disparaîtra, mais une SNV aura été créée (figure 2.5).

Indels

Le second type de mutations regroupe les courtes insertions et délétions (abrégé «indels»). Cela correspond à l'insertion d'une ou de plusieurs paires de nucléotides. Les indels peuvent survenir spontanément lors de la réplication de l'ADN. Ils sont causés, par exemple, par un phénomène appelé glissement réplicatif. Lors de la réplication, un des brins de nucléotides peut former une boucle, et la réplication peut continuer après la boucle. Si la boucle se situe sur le brin parent, une délétion se produit. Au contraire, si la boucle se situe sur le brin enfant, une insertion se produit (figure 2.6).

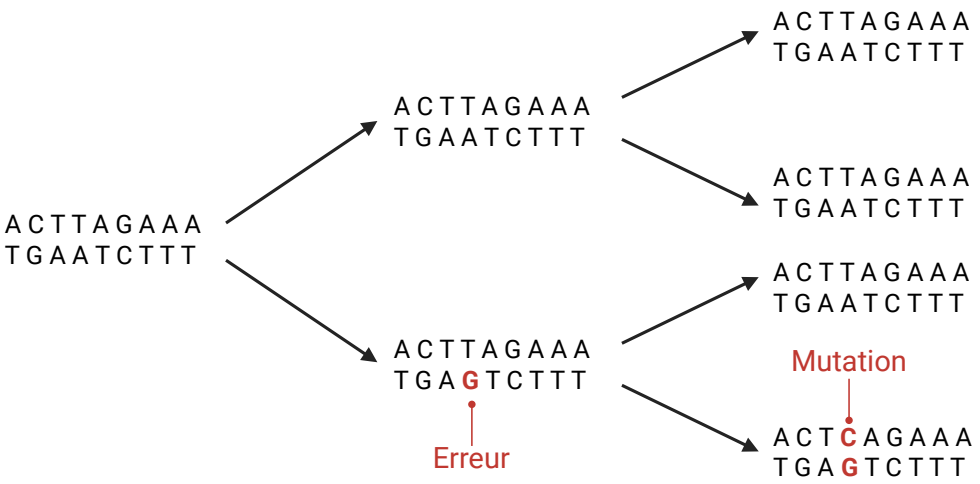


FIGURE 2.5 – Mutation lors de la réplication de l'ADN

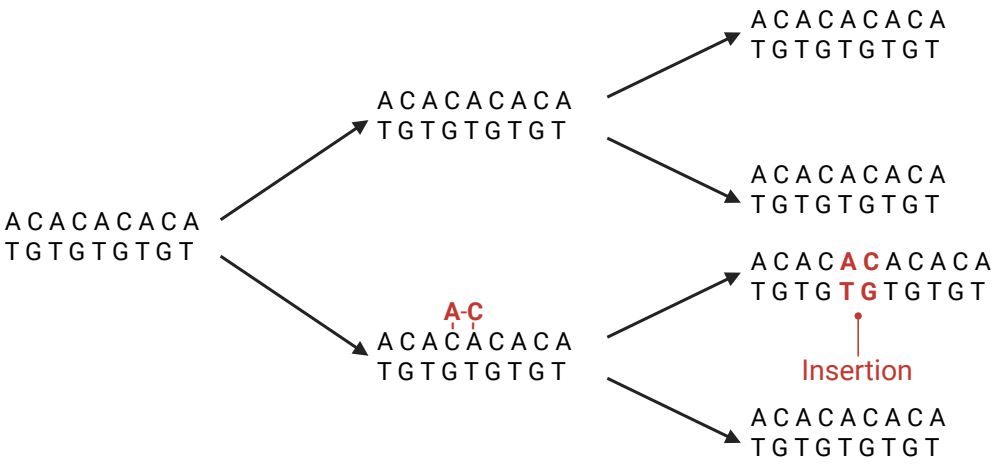


FIGURE 2.6 – Insertion par glissement lors de la réplication de l'ADN

Conséquences des mutations

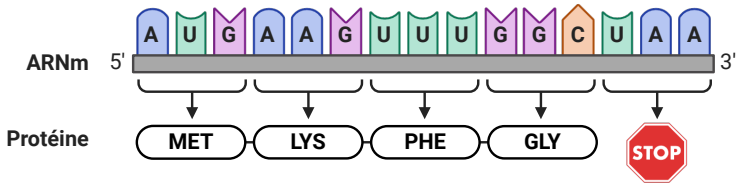
Les conséquences d'une mutation dépendent tout d'abord de la partie du génome où se situe la mutation. Dans le cas où la mutation se produit dans une région codante du génome, les protéines produites peuvent être affectées. Dans les régions codantes, on distingue deux familles de mutations :

- Les mutations synonymes, où la mutation de la séquence d'ADN ne modifie pas la séquence d'acides aminés. Cela est dû au fait que plusieurs codons peuvent désigner le même acide aminé, on parle d'ailleurs aussi de codons synonymes.
- Les mutations non synonymes, où la mutation de la séquence d'ADN entraîne aussi une modification de la séquence d'acides aminés. On distingue plusieurs sous catégories :
  - **Les mutations faux-sens** qui sont des changements dans la séquence d'acide aminé
  - **Les mutations non-sens** qui correspondent à l'introduction de codons de terminaison qui terminent la traduction et réduisent ainsi la taille de la protéine, pouvant la rendre non fonctionnelle.
  - **Les mutations non-stop** qui correspondent à une mutation du codon de terminaison (par exemple son remplacement pas un acide aminé), ce qui a pour effet une traduction continuant au-delà du codon de terminaison, et causant une traduction d'une protéine plus longue qu'initialement prévu.

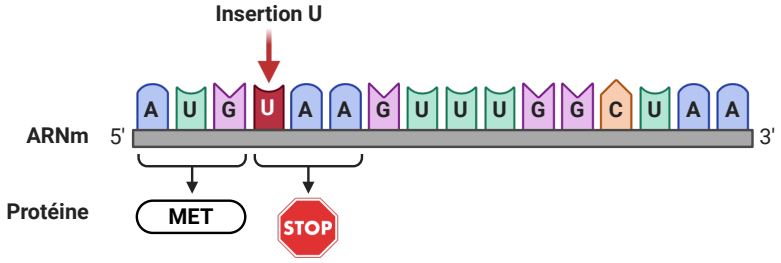
On peut aussi catégoriser les mutations selon leur effet sur l'organisme. Une mutation silencieuse est une mutation qui n'a pas effet sur le phénotype de l'organisme (c'est-à-dire l'ensemble de ses caractéristiques observables). Une mutation synonyme peut être silencieuse, mais elle peut aussi dans certains cas affecter le phénotype [101].

Une mutation peut aussi entraîner l'inactivation totale d'un gène, si le phénotype n'est pas impacté on peut conclure que le gène n'est pas essentiel. Certains gènes sont en effet dupliqués ou leur fonction est dupliquée sur un autre gène.

Concernant les indels, ils peuvent aboutir à un décalage du cadre de lecture. Par exemple l'insertion d'une paire de bases entraîne le décalage de tous les codons (figure 2.7), une séquence d'acides aminés complètement différente et altère le codon de terminaison. Cela peut entraîner un polypeptide anormalement court ou anormalement long, et peut inactiver le gène.



(a) État initial



(b) État après insertion

FIGURE 2.7 – Décalage du cadre de lecture

2.2.3 variations structurales

Les variations structurales correspondent à des modifications larges du génome. Les répétitions dans les génomes constituent des points chauds pour les variations structurales, à savoir les inversions, les délétions, les duplications et les translocations. En général, les répétitions directes stimulent les délétions, les duplications et les translocations, tandis que les répétitions inversées stimulent les inversions [248].

Recombinaison homologue

La recombinaison homologue constitue un mécanisme moléculaire fondamental, essentiel à la fois pour la diversification génétique et la maintenance de l'intégrité génomique. Ce processus implique l'échange de séquences d'ADN entre deux molécules d'ADN qui présentent une homologie de séquence étendue. La recombinaison homologue est également reconnue pour son rôle prépondérant dans la réparation des cassures double-brin de l'ADN, un type de lésion particulièrement préjudiciable.

Les modèles de Holliday et Meselson-Radding, élaborés dans les années 1960 et 1970, ont fourni les premières représentations conceptuelles de la recombinaison homologue, mettant en lumière la formation de structures intermédiaires hétéroduplexes. Les structures hétéroduplexes correspondent à des molécules d'ADN partiellement double brin, où une section ne forme pas de paire de bases. Ces modèles ont été complétés et raffinés par la suite, notamment avec l'introduction du modèle de cassure double-brin, qui propose que la recombinaison homologue soit initiée par une rupture complète des deux brins d'ADN, suivie d'une série d'étapes enzymatiques orchestrant la réparation et l'échange génétique.

Élément transposable

La transposition résulte en le déplacement d'une séquence d'ADN (appelée «transposon») d'une position à une autre dans le génome. On en distingue deux catégories (figure 2.8) :

— Les transposons qui ont une transposition réplicative. Le transposon reste en place et est copié ailleurs dans le génome, c'est un mécanisme de copier-coller.
— Les transposons qui ont une transposition conservative. Le transposon se retire de sa position initiale pour se déplacer ailleurs dans le génome, c'est un mécanisme de couper-coller.

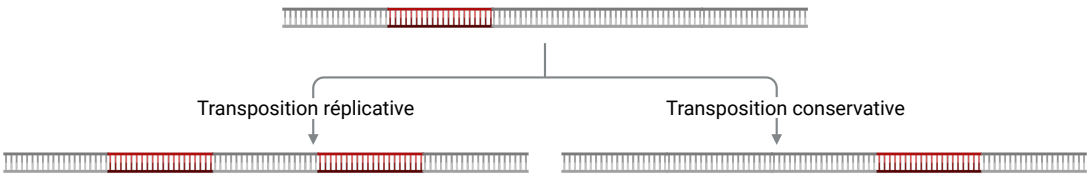


FIGURE 2.8 – Transposition réplicative et conservative

Les éléments transposables les plus simples des bactéries sont les séquences d'insertion (ISs) (figure 2.9). Il s'agit d'une courte séquence d'ADN (entre 750 et 1600 paires de base), encadrée des deux côtés par des répétitions inversées (IRs), et qui contient le gène codant une enzyme requise pour la mobilité de l'IS, la transposase (figure 2.10). Cette protéine reconnaît les IRs qui représentent le début et la fin de IS et permet sa transposition. Par la façon dont la séquence est insérée, des répétitions directes sont créées de part et d'autre de la séquence d'insertion. L'insertion consiste à faire des coupures décalées dans l'ADN cible, à joindre le transposon aux extrémités simple brin saillantes, et à combler les nucléotides manquants. Le comblement des extrémités décalées expliquent l'apparition des répétitions directes de l'ADN cible au site d'insertion. Le décalage entre les coupures

sur les deux brins détermine la longueur des répétitions directes. Ainsi, la répétition cible caractéristique de chaque transposon reflète la géométrie de l'enzyme impliquée dans la coupure de l'ADN cible [152].

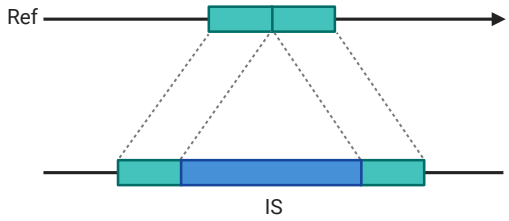


FIGURE 2.9 – Insertion d'un IS dans le génome

L'insertion cause donc l'introduction d'une séquence d'ADN à un endroit du génome. Cette insertion peut interrompre un gène et causer des changements dans son expression, voire son inactivation. Par ailleurs, l'excision incorrecte de l'IS peut laisser une partie de la séquence en place, ce qui résulte en une insertion, ou à l'inverse supprimer une partie de l'ADN, ce qui résulte en une délétion [305].

Types de variations structurales

Délétions Une délétion correspond à la perte d'une section du génome (figure 2.11). Les délétions au sens de variations structurales correspondent à de larges suppressions de sections de génome. La taille des régions supprimées peut aller de quelques centaines de paires de bases à plusieurs milliers. Les délétions larges peuvent être causées par exemple par recombinaison ou lors de la réparation incorrecte d'une cassure.

Inversions Les inversions correspondent à une variation structurale de la séquence d'ADN où une section du génome remplace la section initiale, mais dont l'ordre des nucléotides est inversé (figure 2.12).

Duplications Les duplications correspondent à une variation structurale de la séquence où une section du génome se duplique. Quand le segment dupliqué est adjacent au segment original, on parle de duplication en tandem, à mettre en opposition aux duplications déplacées qui sont à une certaine distance du segment d'origine. Le segment dupliqué peut être inversé par rapport au segment original, dans ce cas on parle de duplication inverse ou en miroir.

Comme exemple de duplication, on trouve les répétitions directes créées par les séquences d'insertions, ou les répétitions inversées qui les entourent. Les duplications peuvent aussi être très larges, on a par exemple chez *Mycobacterium tuberculosis* variant *bovis* une duplication de 100 000 nucléotides[26].

Translocation La translocation correspond au déplacement d'une région du génome à une autre position dans le génome (figure 2.13).

2.3 Méthodes d'étude des génomes

Afin d'étudier la séquence d'ADN d'un génome, il est nécessaire de déterminer sa séquence de nucléotides, ou au moins des fragments de cette séquence. Pour cela, on utilise différentes méthodes de biologie moléculaire. Chaque méthode offre un niveau de visibilité différent sur le génome, et est plus ou moins applicable à large échelle. Là où l'étude *in-silico* des génomes se fait majoritairement à l'aide de données de Séquençage

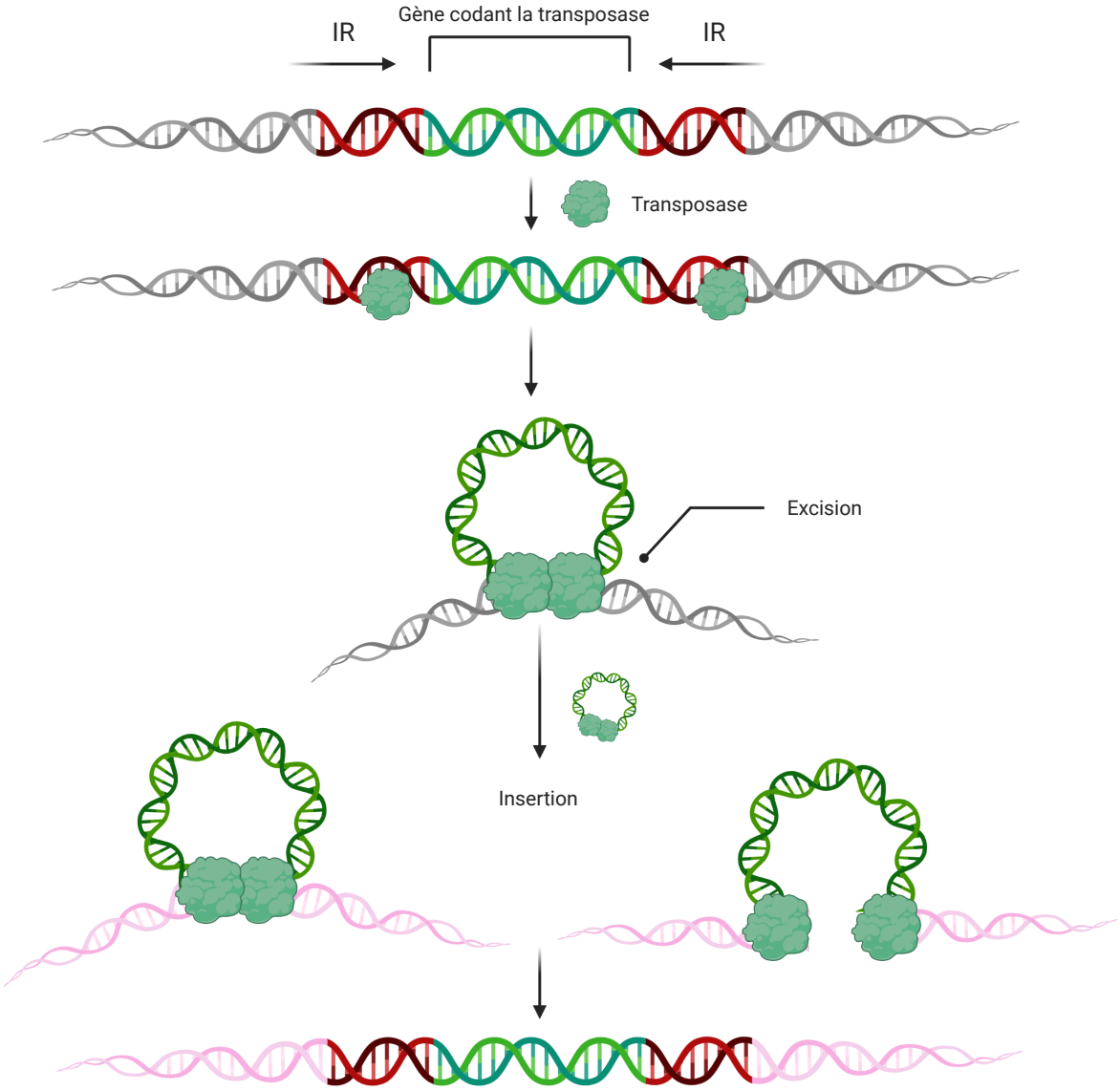


FIGURE 2.10 – Transposition d'une séquence d'insertion par la transposase

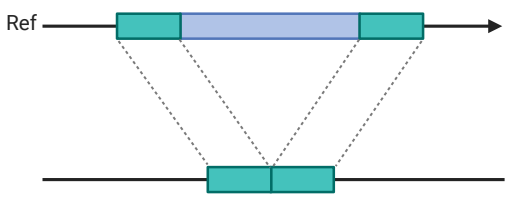


FIGURE 2.11 – Délétion d'une région d'un génome

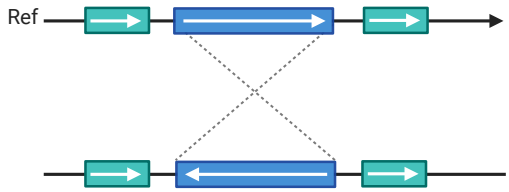


FIGURE 2.12 – Inversion d'une région d'un génome

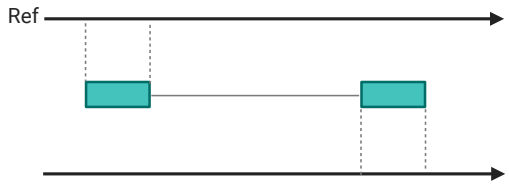
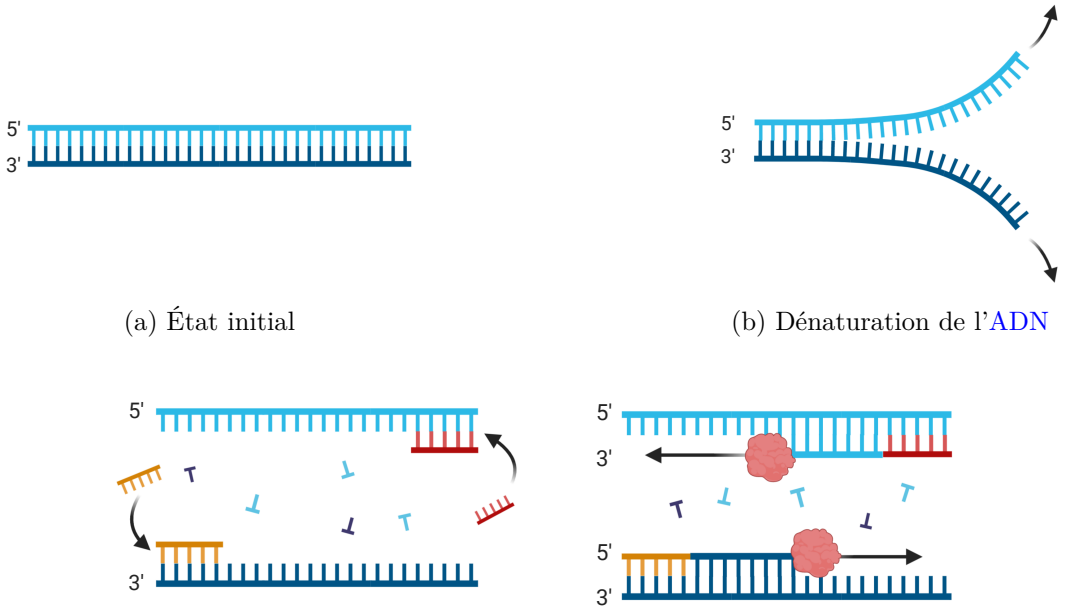


FIGURE 2.13 – Translocation d'une région d'un génome

de nouvelle génération (NGS), de nombreuses données présentes dans la littérature sont issues d'autres méthodes.

2.3.1 PCR

Le principe de réaction en chaîne par polymérase (PCR) a été inventé par Kary Mullis en 1983. Il permet l'amplification *in vitro* de l'ADN, c'est-à-dire la synthèse rapide de milliards de copies d'un fragment d'ADN précis depuis un ensemble complexe d'ADN. La PCR est ainsi utile pour, par exemple, aider à la détection de fragments d'ADN spécifiques.



(a) État initial

(b) Dénaturation de l'ADN

(c) Hybridation

(d) Extension

FIGURE 2.14 – Réaction en chaîne par polymérase

La préparation de la PCR commence par la synthèse chimique d'oligonucléotides, des chaînes simples de nucléotides complémentaires des séquences encadrant la séquence à

amplifier. Les oligonucléotides sont produits avec un synthétiseur et leur taille est généralement comprise entre 15 et 30 nucléotides. Elles vont servir d'amorce pour la synthèse de l'ADN lors de la PCR.

Le milieu réactionnel comprend ainsi les oligonucléotides servant d'amorces, l'ADN contenant la séquence à amplifier, une ADN polymérase thermostable (c'est-à-dire étant stable à haute température, par exemple la Taq polymérase est souvent utilisée) et les 4 désoxyribonucléosides triphosphates (appelés dNTPs regroupant dATP, dCTP, dGTP, dTTp). Une succession de réactions répétées, appelées cycles, sont exécutées par un thermocycleur[298]. Chaque cycle comprend trois étapes.

— La première est la dénaturation (figure 2.14b) de l'ADN contenant la séquence à amplifier, c'est-à-dire la séparation de l'ADN en deux simples brins par la rupture des liaisons hydrogène. Cette étape est réalisée en augmentant la température à 95°.

— Dans la seconde étape, la température est abaissée entre 50 et 68°C ce qui permet aux amorces de se lier sur chaque brin autour de la séquence visée par le fait que leur séquence est complémentaire. Ce processus est appelé hybridation ou annelage (*annealing* en anglais) (figure 2.14c).

— Dans la dernière étape, la température est relevée entre 68 et 72°C, ce qui permet à l'ADN polymérase d'étendre les amorces et synthétiser la copie de la séquence d'ADN à amplifier. Ce processus est appelé extension (figure 2.14d). À noter que les températures sont indicatives et varient en fonction des protocoles.

À chaque cycle ces étapes sont répétées, la séquence d'ADN visée est ainsi copiée de manière exponentielle. Ces copies sont appelées amplicons.

La méthode la plus commune pour visualiser le produit de la PCR est l'électrophorèse sur gel de polyacrylamide ou d'agarose, suivie d'une coloration. Quand des molécules d'ADN sont placées au pôle négatif d'un champ électrique, elles migrent vers le pôle positif. La migration se fait en fonction de la masse moléculaire des fragments (plus la masse est faible, plus le déplacement est rapide) et de la concentration en gel [304].

Quand la PCR est utilisée pour détecter la présence ou l'absence d'un produit spécifique on parle de PCR qualitative. La PCR quantitative (aussi appelée temps réel ou réaction en chaîne par polymérase en temps réel (qRT-PCR)) va plus loin en donnant une estimation de la quantité originale de la séquence amplifiée. On mesure la quantité de produit formée après les cycles, avec un marqueur fluorescent ou des sondes. [154]

2.3.2 Séquençage par la méthode de Sanger

La technique du séquençage est utilisée pour identifier l'ordre des nucléotides dans l'ADN. À partir d'un échantillon d'ADN, on peut ainsi déterminer la séquence de nucléotides qui le compose. La méthode de Sanger, aussi connue sous le nom de séquençage par terminaison de chaîne, est une procédure développée en 1977 par Frederick Sanger et son équipe. Bien qu'elle puisse être effectuée manuellement, elle est souvent automatisée grâce à l'utilisation d'un séquenceur.

Ce séquençage repose sur le principe qu'une molécule d'ADN simple brin, dont la taille diffère ne serait-ce que d'une seule nucléotide, peut être séparée des modèles d'autre taille par électrophorèse sur gel de polyacrylamide. Par exemple, lorsqu'on réalise l'électrophorèse dans un tube capillaire mesurant entre 50 et 80 cm de long et ayant un diamètre interne de 0,1 mm, il devient possible de séparer un ensemble de molécules de différentes tailles, allant jusqu'à 1500 nucléotides. Dans ce processus, les molécules simple brin sont réparties successivement d'une extrémité à l'autre du tube capillaire [27].

De manière similaire à la PCR, le milieu réactionnel comprend les oligonucléotides servant d'amorces, l'ADN contenant la séquence à séquencer, un ADN polymérase thermostable (par exemple la Taq polymérase) et les 4 désoxyribonucléosides triphosphates (appelés dNTPs regroupant dATP, dCTP, dGTP, dTTp). La réaction de synthèse devrait normalement produire des séquences de milliers de nucléotiques, mais ce n'est pas le cas ici.

En effet, on introduit en plus de petites quantités des quatre triphosphates de didésoxynucléotides (ddNTPs : ddATP, ddCTP, ddGTP et ddTTP). Chacun de ces didésoxynucléotides est marqué par un marqueur fluorescent distinct (figure 2.15a). La particularité de ces ddNTPs est qu'ils bloquent l'élongation du brin d'ADN, car il leur manque le groupe hydroxyle en position 3' nécessaire pour établir une liaison avec le nucléotide suivant. La proportion faible de ddNTPs par rapport aux dNTPs fait que l'élongation peut se faire sur une partie des brins avant d'être stoppée. L'extension étant toujours de $5' \rightarrow 3'$, les brins produits commencent de l'amorce jusqu'à être arrêtés par les ddNTPs. En résumé, on suit le même mécanisme que la PCR où la séquence est amplifiée, sauf que lors de l'extension (figure 2.15b) la polymérase est parfois bloquée dans son action (figure 2.15c).

Au final, on obtient un ensemble de molécules d'ADN simple brin, qui correspondent à l'extension progressive, à partir de l'amorce, de la molécule d'ADN à séquencer (figure 2.15d). L'électrophorèse sur gel permet ensuite de séparer les molécules en fonction de leur taille, de la plus courte à la plus longue. Suite à leur séparation, les molécules sont analysées par un détecteur de fluorescence capable de distinguer les marqueurs liés aux didésoxynucléotides. Ainsi, le détecteur est en mesure d'identifier si chaque molécule se termine par A, C, G, ou T. La séquence obtenue est sauvegardée sous forme de fichier, ce qui permet de réaliser des analyses *in silico* dessus (figure 2.15f).

2.3.3 Séquençage NGS (Illumina)

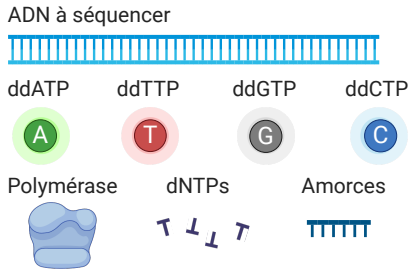
La méthode de Sanger présente des limitations significatives, notamment sa capacité à séquencer uniquement un fragment d'ADN à la fois. Cette contrainte limite considérablement le volume de séquençage et rend l'analyse de grands génomes ou de multiples échantillons à la fois coûteuse et chronophage. Face à ces limitations, les technologies de Séquençage de nouvelle génération (NGS) ont émergé, dont la plateforme commerciale Illumina. Contrairement à la méthode de Sanger, le NGS permet le séquençage parallèle massif de millions de fragments d'ADN simultanément par exécution.

Dans une première phase appelée «préparation de la librairie» (figure 2.16a), la molécule d'ADN à séquencer est fragmentée en courtes sections appelées «inserts». L'insert fait en général quelques centaines de bases. Lors de la préparation, des morceaux d'ADN appelés «adaptateurs» sont attachés aux deux extrémités de chaque insert. La molécule d'ADN ainsi formée, comprenant l'insert et les deux adaptateurs, est appelée «fragment». Il y a deux types d'adaptateurs, l'un se place au début des séquences ($5'$) et l'autre à la fin ($3'$). À la fin de la première phase, les fragments sont dénaturés. Les deux brins d'ADN se séparent en deux polynucléotides.

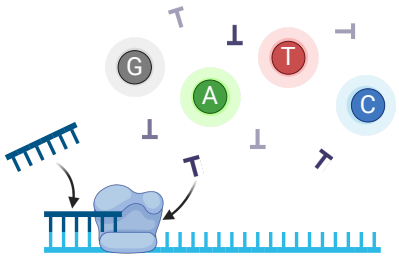
La seconde étape, dite d'«hybridation», consiste à fixer ces fragments sur une lame de verre appelée «*flow cell*» (figure 2.16b). En réalité les fragments ne sont pas fixés mais copiés sur la *flow cell*. La *flow cell* contient des oligonucléotides complémentaires aux adaptateurs. Une fois l'adaptateur fixé à l'oligonucléotide complémentaire, l'ADN polymérase crée le complément du fragment hybridisé. Il s'ensuit une dénaturation et le fragment ayant servi de modèle est ainsi supprimé. À la fin de cette seconde étape, tous les fragments sont donc fixés à la *flow cell*.

À ce stade, la *flow cell* contient des fragments dispersés sur une fraction de ses oligonucléotides. Ces fragments sont difficilement détectables, donc une amplification est nécessaire. La troisième étape est une amplification par PCR en «pont» (figure 2.16b). L'adaptateur resté non fixé à la *flow cell* dans l'étape précédente crée un pont avec un oligonucléotide complémentaire proche sur la *flow cell*. L'ADN polymérase crée le complément du fragment hybridisé et il s'en suit une dénaturation, le fragment ayant servi de modèle est ainsi supprimé. Les fragments hybridisés sur un des deux types d'oligonucléotides sont ensuite détachés, laissant ainsi tous les fragments de la *flow cell* orientés dans le même sens. A la fin de cette étape, des clusters de fragments, tous orientés dans le même sens, sont ainsi formés.

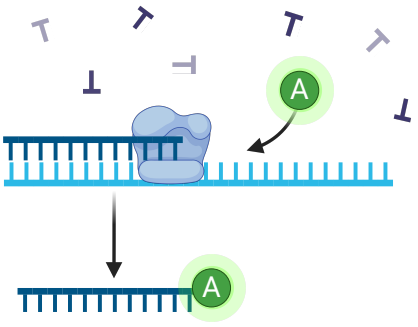
La dernière étape est le séquençage (figure 2.16c). Des amorces se lient aux adapteurs.



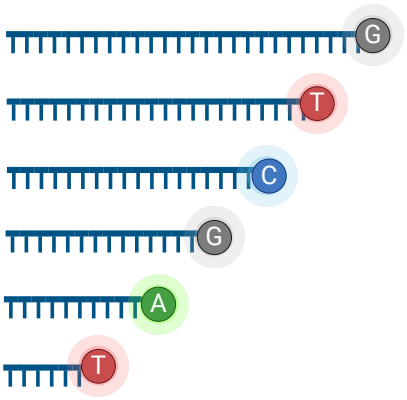
(a) Milieu réactionnel au début du séquençage



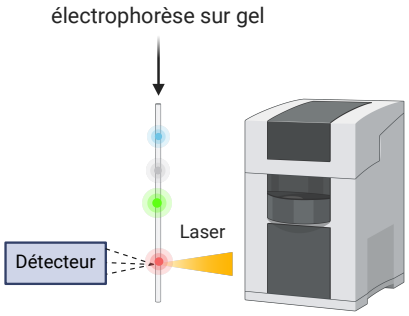
(b) Hybridation et extension de l'ADN



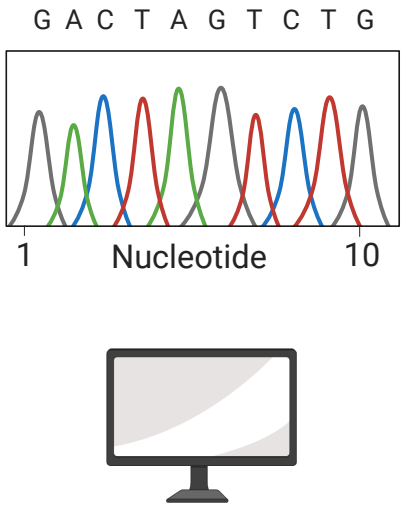
(c) Lors de l'extension, ddNTP peut parfois être placé au lieu d'un dNTP, cela a pour conséquence de terminer la chaîne



(d) On retrouve après la réaction un ensemble de molécules simple brin d'ADN avec en fin de chaîne des didésoxynucléotides marquées par un marqueur fluorescent



(e) L'électrophorèse sur gel permet la séparation des molécules en fonction de leur taille. On détecte ensuite les marqueurs fluorescents qui permettent de connaitre le dernier nucléotide de chaque groupe de molécules d'ADN

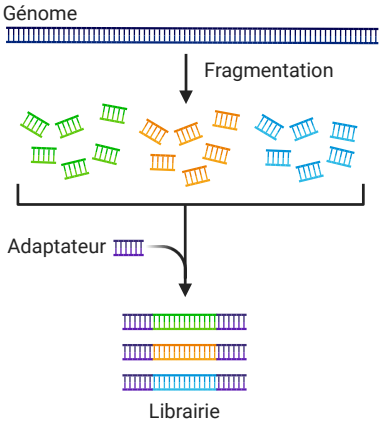


(f) L'analyse de la quantité de marqueurs par rapport à la position sur le gel permet d'obtenir la séquence

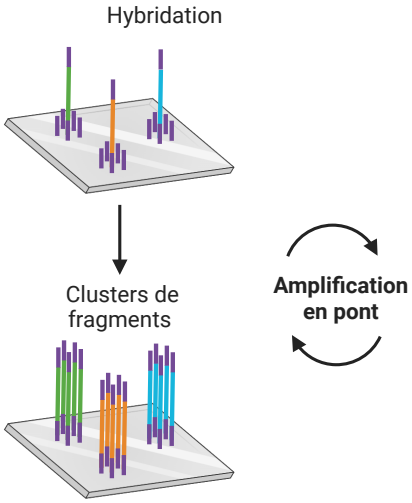
FIGURE 2.15 – Séquençage par la méthode de Sanger

L'ADN polymérase ajoute ensuite un nucléotide marqué par un marqueur fluorescent pour créer le complément du fragment. Les marqueurs sont excités par un laser sur la *flow cell*, ce qui permet de visualiser quel nucléotide s'est attaché au fragment dans chaque cluster. Le marqueur est ensuite supprimé. Le cycle se répète ainsi de nombreuses fois dans un processus massivement parallèle.

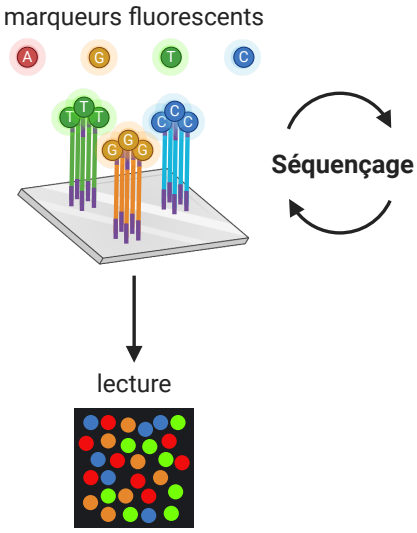
Le séquençage single-end et le séquençage paired-end sont deux méthodes utilisées dans les systèmes de séquençage Illumina, chacune ayant des avantages et des applications distincts. Le séquençage single-end lit l'ADN à partir d'une seule extrémité du fragment d'ADN, offrant une approche simple et économique. Cette méthode fournit rapidement de grandes quantités de données, mais peut limiter la capacité à détecter certaines caractéristiques génomiques telles que les indels ou les réarrangements complexes. En revanche, le séquençage paired-end lit les deux extrémités du fragment d'ADN, doublant efficacement la quantité de données obtenues à partir de la même quantité d'ADN utilisée dans le séquençage single-end. Cette approche, non seulement améliore la qualité, mais améliore également la détection des réarrangements génomiques et des éléments répétitifs.



(a) Préparation de la librairie



(b) Amplification en pont



(c) Cycles de séquençage

FIGURE 2.16 – Séquençage NGS (Illumina)

2.3.4 Séquençage de troisième génération

Les séquences issues de la seconde génération de séquençage étaient plus courtes par rapport au séquençage de Sanger, ce qui empêchait de séquencer précisément des répétitions plus longues que la taille d'insert.

Au cours de la dernière décennie, une troisième génération de technologies de séquençage, «*long read sequencing*», incluant les technologies de Pacific Biosciences (PacBio) et d'Oxford Nanopore Technologies (ONT), a permis aux chercheurs de générer des séquences de haute qualité, sans fragmentation des reads. Les technologies long read sequencing peuvent produire des lectures allant jusqu'à des dizaines de milliers de paires de bases. Par exemple, les lectures continues longues obtenues avec une machine PacBio Sequel II peuvent atteindre une longueur N50 brute de 30 à 60 kb et une précision de 87 à 92%. N50 étant la longueur du plus court contig pour lequel les contigs de longueur supérieure ou égale couvrent au moins 50 % du génome. Les séquenceurs ONT MinIon/GridION peuvent produire de longues et ultra-longues lectures avec un N50 de 10 à 60 et de 100 à 200 kb, respectivement, avec une précision de 87 à 98% [222]. Ce type de technologie présentant un taux d'erreur important, un séquençage de seconde génération est donc parfois utilisé en complément.

La technologie PacBio fonctionne à l'aide d'un guide d'ondes en mode zéro, qui est un dispositif optique permettant d'observer la copie d'une unique molécule d'ADN. Cette technologie ne nécessite pas d'interruption entre chaque lecture comme sur la technologie Illumina, la lecture s'effectue en continu, d'où le terme parfois utilisé de séquençage en temps réel [27].

La technologie Nanopore fonctionne avec le passage de la molécule d'ADN à travers des nanopores (trou avec un diamètre de l'ordre du nanomètre). Un courant électrique traverse le nanopore lorsque ce dernier est obstrué par un nucléotide de la molécule d'ADN. La quantité de courant qui peut passer dans le nanopore diffère en fonction du nucléotide (T, A, G ou C) ce qui permet de déterminer la séquence (figure 2.17).

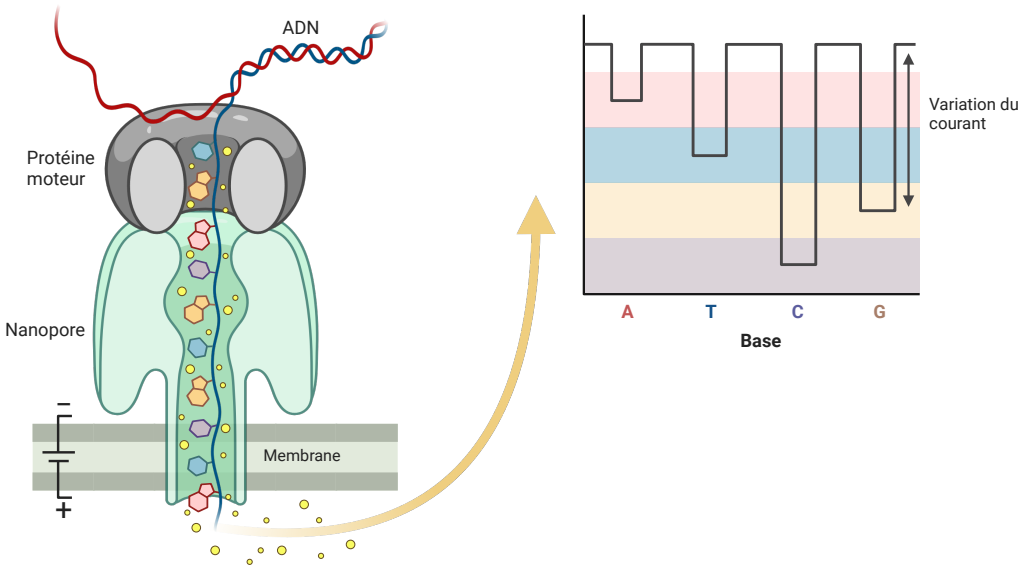


FIGURE 2.17 – Séquençage avec la technologie nanopore

2.4 La tuberculose

Un pathogène est défini comme un organisme causant une maladie à son hôte, appelée maladie infectieuse. La gravité des symptômes de la maladie est appelée virulence. Les pa-

thogènes sont très divers et comprennent des virus et des bactéries, ainsi que des eucaryotes unicellulaires et multicellulaires. Chaque organisme vivant est affecté par des pathogènes, y compris les bactéries, qui sont ciblées par des virus spécialisés appelés phages. Les pathogènes peuvent causer des maladies de nombreuses manières. Bien souvent cela passe par la production de toxines qui causent des dommages aux tissus ou aux cellules de l'hôte. Les toxines bactériennes figurent parmi les poisons les plus mortels connus et comprennent par exemples le tétanos, l'anthrax ou la toxine botulique [12].

Depuis les années 1960, les progrès en matière d'analyse génomique et génétique ont permis d'importantes avancées quant à la classification des micro-organismes. Les travaux de comparaison de l'ARN ribosomique commencés par Carl Woese dans les années 1970 ont été décisifs dans la démonstration de l'existence de deux domaines distincts que sont les bactéries et archées, là où se trouvaient les cellules prokaryotes [304]. C'est en 1977 que Woese et Fox montrèrent [306], par la comparaison de la séquence nucléotidique de l'ARN ribosomique 16S de différentes cellules procaryotes, une relation d'évolution parmi des organismes très différents. [227]

Parmi les agents infectieux, les bactéries sont les plus petits organismes capables d'exister de manière indépendante [229]. Elles sont morphologiquement très diverses, mais sont majoritairement des bacilles (forme allongée en bâtonnet) ou des cocci (forme sphérique). Elles peuvent exister seules ou être assemblées dans des arrangements caractéristiques utiles pour leur identification [304]. Leur cytoplasme contient des ribosomes ainsi qu'en général un unique chromosome. Bien qu'elles n'aient pas de noyau, tous les éléments chimiques de l'acide nucléique et pour la synthèse protéique sont présents [229].

La tuberculose est une maladie infectieuse causée par des bactéries membres du complexe *Mycobacterium tuberculosis* complex (MTBC), un groupe de bacilles étroitement liés (avec une séquence d'ADN commune à plus de 99 %). La tuberculose affecte la population mondiale depuis l'antiquité, et a eu un impact majeur sur la population mondiale, se propageant par la méconnaissance de la maladie, la pauvreté et le manque d'hygiène. L'homme reste le principal porteur, même si un certain nombre d'animaux peuvent être infectés. La transmission est majoritairement due à l'inhalation de gouttelettes contenant la mycobactérie. Mais elle peut, dans des cas plus rares, être transmise par voie gastro-intestinale après ingestion de lait contaminé (peu probable de par la pasteurisation du lait), ou par une plaie sur la peau [229].

Le terme «tuberculose» a été introduit pour la première fois par Johann Lukas Schönlein en 1839. Ce terme dérive de «tubercule», qui désigne la lésion caractéristique de cette maladie. Utilisé depuis le 17e siècle, «tubercule» vient du latin «tuber», qui signifie «excroissance».

2.4.1 Histoire

Préhistoire Actuellement, il n'existe pas de consensus sur l'âge du plus récent ancêtre commun du MTBC [92]. Une étude a estimé que cet ancêtre commun vivait il y a environ 70 000 ans [48]. En revanche, deux autres recherches, utilisant de l'ADN ancien du MTBC extrait de momies hongroises vieilles d'environ 200 ans et de restes humains péruviens datant d'environ 1 000 ans, ont proposé une estimation de moins de 6 000 ans [146, 20].

Ces estimations plus récentes sont en contradiction avec les preuves trouvées dans des restes humains vieux d'environ 11 000 ans en Syrie [11] et d'environ 9 000 ans en Israël [116], ainsi que dans les restes d'un bison au Wyoming datés d'environ 17 000 ans [158]. Cependant, l'authenticité de certaines de ces découvertes antérieures a été mise en question.

Antiquité On retrouve des preuves de l'existence de la maladie à l'époque de l'antiquité par l'analyse des ossements. Des momies égyptiennes datant de 2400 av. J.-C. montrent des déformations du squelette typiques de la tuberculose [14], observations confirmées par analyses microscopiques [316]. Le texte juridique le plus ancien du monde, écrit par

le monarque babylonien Hammurabi entre 1948 et 1905 av. J.-C. et gravé sur une stèle de pierre, mentionne une maladie pulmonaire chronique, qui était très probablement la tuberculose [117]. D'autres documents écrits sont associés à l'hébraïsme. Le mot hébreu ancien schachepheth est utilisé dans les livres bibliques du Deutéronome et du Lévitique pour décrire la tuberculose [14].

Au temps du médecin et philosophe Hippocrate, apparaît dans la littérature grecque le terme de «phtisie», ou consomption (affaiblissement et maigreur extrêmes). Hippocrate a détaillé les symptômes de la consomption dans son livre 1 de «Epidemiai», caractérisée par des fièvres continues avec des frissons, des sueurs, des membres froids, et des troubles intestinaux. Il décrit que les patients présentaient également des urines anormales, des crachats difficiles à expulser, et des gorges douloureuses avec inflammation. Cette maladie entraînait une détérioration rapide, une perte d'appétit et, souvent, un état délirant avant la mort [119].

Les patients atteints de la maladie étaient traités dans les temples avec de la nourriture saine, du lait et de l'exercice physique. Hippocrate a été le premier à observer des lésions semblables à des tubercules dans les tissus de divers animaux, mais rien de semblable n'a été décrit chez l'homme à cette époque car les autopsies humaines n'étaient pas réalisées. L'école hippocratique pensait que la phtisie pulmonaire était héréditaire plutôt qu'infectieuse. En revanche, Aristote (384 à 322 av. J.-C) croyait que la phtisie était contagieuse, contrairement à l'opinion populaire de l'époque qui la considérait comme héréditaire. Aristote a aussi décrit le scrofule sur la peau chez les porcs ayant la phtisie, qui sont des fistules purulentes localisées sur les ganglions lymphatiques du cou [117].

Claude Galien, un médecin grec au temps de l'empereur romain Marc Aurèle en 174 après J.-C., a décrit la phtisie comme accompagnée de fièvre, de sueurs et de toux de crachats teintés de sang, et a trouvé les tubercules dans les poumons phtisiques qu'il appelait phûma. Il considérait cela comme infectieux et mettait en garde contre le contact étroit avec les personnes atteintes de la maladie [89].

Renaissance En 1679, Sylvius de la Boë a été probablement le premier à utiliser le terme "tubercules" pour désigner une affection pulmonaire, qu'il nommait "tubercula glandulosa", et a observé leur évolution vers des complications graves. Il a également fait le lien entre la phtisie et le scrofule. Plus tard, l'association entre la phtisie, les tubercules pulmonaires et la scrofule a été également discutée par Richard Morton, qui a émis l'hypothèse que la phtisie pouvait être héréditaire, et potentiellement transmissible par contact proche [89].

Quelques années plus tard, Jean-Jacques Manget, un médecin genevois, observa lors d'une autopsie de multiples petits nodules phtisiques dans les poumons et les organes qui ressemblaient à des graines de millet, plus tard appelés "tuberculose miliaire" [89].

Époque moderne Le terme de consomption a été utilisé par les non-initiés pour décrire la phtisie aux XVIIe et XVIIIe siècles, et les deux termes étaient utilisés jusqu'au milieu du XIXe siècle, lorsque le terme tuberculose a été inventé par Johann Lukas Schönlein et plus tard utilisé par Hermann Brehmer, Jean Antoine Villemin et Robert Koch.

René Laennec, fut l'un des médecins les plus importants à se consacrer à la maladie dont il décédera à l'âge de 45 ans, victime de la tuberculose qu'il avait contractée lors de ses recherches sur des patients contagieux et des corps infectés [280]. Au cours de l'année 1816, pendant qu'il exerçait à l'Hôpital Necker à Paris, il a conçu le premier stéthoscope en improvisant avec son carnet roulé, avant de fabriquer un modèle plus abouti sous la forme d'un cylindre en bois creux.Il rédigea des descriptions précises de diverses pathologies thoraciques telles que la bronchiectasie, la pneumonie, la pleurésie, l'emphysème et la tuberculose. Grâce à son stéthoscope, Laennec a pu détecter chez les patients atteints de tuberculose des anomalies telles que la consolidation, la pleurésie et la cavitation pulmonaire.

Laennec caractérisa avec précision les tubercules pulmonaires et extrapulmonaires,

révélant leur rôle initial dans le développement de la phtisie. Il décrivit l'évolution des tubercules, depuis leur apparition initiale dans le poumon sous une forme granulaire, comparée à celle de graines de millet, jusqu'à leur transformation en structures plus volumineuses renfermant une substance caseuse, leur rupture en pus, et enfin la formation d'abcès et d'empyèmes. De plus, Laennec décrivit les manifestations de la tuberculose hors des poumons, dans des organes tels que les intestins, le foie, les méninges, ainsi que dans les vertèbres. Il a mis en lumière l'infection tuberculeuse vertébrale provoquant l'effondrement des vertèbres et la paralysie de la moelle épinière, pathologie préalablement décrite par Percivall Pott, chirurgien britannique en 1779, connue sous le nom de maladie de Pott [89].

En 1869, Jean Antoine Villemin démontra que la maladie était contagieuse, en menant une expérience dans laquelle une petite quantité de liquide purulent provenant d'une cavité tuberculeuse prélevé sur un cadavre humain était injectée dans des lapins de laboratoire, qui devenaient alors infectés. Du sang ou des crachats de lapins tuberculeux, injectés dans d'autres animaux de laboratoire produisaient une tuberculose, alors que le transfert analogue de tissu cancéreux ou fibrotique n'a eu aucun effet. Il a alors postulé pour la première fois que la maladie est causée par un micro-organisme spécifique, qui doit être présent dans l'air [117, 14].

Laennec et Villemin, considéraient la tuberculose comme complètement incurable. Laennec dit "Presque tous les hommes de l'art pensent aujourd'hui que l'affection tuberculeuse est, comme les affections cancéreuses, absolument incurable." [117].

L'histoire de la tuberculose a connu un tournant décisif le 24 mars 1882, lorsque Hermann Heinrich Robert Koch, élève de Henle, a révélé au monde sa découverte majeure lors de sa présentation intitulée "Die Ätiologie der Tuberkulose" devant la Société de Physiologie à l'hôpital de Charité à Berlin. Au cours de cette conférence, Koch a démontré, pour la première fois, l'existence du bacille responsable de la tuberculose, *Mycobacterium tuberculosis*. Grâce à des techniques de coloration, il a réussi à mettre en évidence des bacilles tuberculeux au sein de tissus tuberculeux. Il parvint aussi à cultiver la bactérie pour l'inoculer à des animaux de laboratoire, et ainsi démontrer que *Mycobacterium tuberculosis* était bien l'agent étiologique. En plus de cette découverte fondamentale, Koch présenta ses postulats, qui définissent encore aujourd'hui les standards pour démontrer l'étiologie des maladies infectieuses. Ces postulats sont parfois mieux connus sous le nom de postulats Koch-Henle [68, 117].

En 1890, Koch développa la tuberculine, un extrait à la glycérine de bacilles tuberculeux morts. La tuberculine fut présentée comme un remède efficace, et a suscité un intérêt mondial poussant d'innombrables patients atteints de tuberculose à se rendre à Berlin dans l'espoir de trouver un traitement. Cependant, il s'est finalement avéré que la tuberculine était inefficace. La tubuculine servira cependant de base pour le développement d'un test de diagnostic de la tuberculose avec le travail de Clemens von Pirquet en 1907 et Charles Mantoux en 1908.

Un vaccin fut développé par Albert Calmette et Camille Guérin, et fut testé pour la première fois en 1921. Cependant ce vaccin a une efficacité limitée malgré avoir été administré à plus de 3.5 milliards de personnes [80].

2.4.2 Infection et transmission

Comme la plupart des infections respiratoires, la tuberculose est principalement transmise par voie aérienne, depuis des sources humaines ou animales, par l'inhalation de fines particules aéroportées. Ces gouttelettes sont de petites particules infectieuses (<6 µm) de sécrétions respiratoires qui sont aérosolisées en toussant, éternuant, parlant ou chantant. On estime que la taille des gouttelettes portant les bacilles de la tuberculose se situe entre 1 et 5 µm [40].

En fonction de la qualité de l'air, ces particules peuvent rester en suspension plusieurs heures. Lorsqu'inhalées par autrui, elles peuvent engendrer une nouvelle contamination.

Le risque de transmission dépend de plusieurs facteurs : le degré de contagiosité du patient, l'environnement dans lequel la transmission a lieu, la durée de l'exposition et la vulnérabilité du nouvel hôte face au pathogène.

Être infecté par le bacille de la tuberculose, *M. tuberculosis*, ne signifie pas systématiquement développer la maladie sous sa forme dite active, qui est contagieuse. En effet, seul un individu immunocompétent sur dix portant le bacille passe à une phase active de la maladie. Le risque est donc beaucoup plus élevé chez les personnes présentant une immunodépression ou infectées par le VIH. Environ 90 % des porteurs ne verront donc jamais la maladie se manifester activement et ne constitueront pas un vecteur de contagion [32].

Quand la tuberculose n'est pas active, on parle de tuberculose latente. L'infection latente est définie par l'Organisation mondiale de la santé comme «Un état de réponse immunitaire persistante à la stimulation par les antigènes de *Mycobacterium tuberculosis* sans preuve d'une tuberculose active cliniquement manifeste.». Comme il n'existe pas de test de référence pour la tuberculose latente, on ne connaît pas avec certitude la part de la population mondiale infectée. On estime cependant que jusqu'à un quart [39] de la population mondiale est infectée par *M. tuberculosis*, et la grande majorité ne présente aucun signe ni symptôme de la maladie. En moyenne, 5 à 10 % des personnes infectées développeront la tuberculose active au cours de leur vie, généralement dans les cinq premières années suivant l'infection initiale [276].

Le passage d'une infection latente à une maladie active peut être rapide ou prendre des années, selon les cas. Ce processus est influencé par plusieurs variables personnelles comme l'état immunologique, l'état de santé général du sujet, le moment de l'infection originelle, la présence de certaines pathologies (telles que le diabète ou le cancer) et le recours à des thérapies immunosuppressives [32].

Il existe d'autres vecteurs d'infection bien qu'ils soient plus rares. L'infection peut faire suite à une introduction directe de *Mycobacterium tuberculosis* dans la peau ou les muqueuses d'un individu sensible suite à un traumatisme ou une blessure. La contamination peut aussi se faire par ingestion de lait contaminé pour le variant *bovis* de *Mycobacterium tuberculosis*.

2.4.3 La maladie

La tuberculose a la capacité d'affecter diverses régions du corps, toutefois, elle est principalement rencontrée dans les poumons, situation désignée sous le terme de tuberculose pulmonaire. Lorsque la maladie se développe autre part que dans les poumons, on parle de tuberculose extrapulmonaire, laquelle peut se présenter simultanément avec la forme pulmonaire de la maladie.

Tuberculose pulmonaire La tuberculose pulmonaire est définie comme la tuberculose du parenchyme pulmonaire et de l'arbre bronchique uniquement. Les caractéristiques cliniques de la tuberculose pulmonaire incluent la toux chronique, la production de crachats, la perte d'appétit, la perte de poids, la fièvre, les sueurs nocturnes et l'hémoptysie (toux ramenant du sang en provenance des voies respiratoires) [168].

On peut distinguer la tuberculose pulmonaire primaire, il s'agit de la première manifestation de la tuberculose après une infection. Dans les pays où la prévalence de la tuberculose est élevée, la tuberculose pulmonaire primaire survient généralement dans l'enfance, mais là où la tuberculose est moins endémique, elle survient également assez souvent chez les adultes. Elle se caractérise par une inflammation granulomateuse locale, généralement en périphérie du poumon (appelée foyer de Ghon), et peut être accompagnée d'une atteinte des ganglions lymphatiques, appelée le complexe de Ghon. L'infection est généralement asymptomatique mais peut se présenter sous la forme d'une infection aiguë des voies respiratoires inférieures. La tuberculose peut ensuite devenir latente chez une personne immunocompétente.

Une tuberculose post-primaire, ou réactivation, peut suivre la tuberculose primaire. Chez les personnes généralement immunocompétentes, il existe une chance à vie de réactivation de la tuberculose dormante de 5 % à 10 %. Une autre cause de tuberculose post-primaire est la réinfection de la personne [168].

Tuberculose extra-pulmonaire La tuberculose extra-pulmonaire est définie comme étant une tuberculose affectant tout autre endroit que les poumons. Une atteinte extra-pulmonaire peut être observée chez plus de 50 % des patients atteints simultanément du SIDA et de la tuberculose. En effet, le risque de tuberculose extrapulmonaire augmente avec l'avancement de l'immunosuppression [99].

La lymphadénite (infection aiguë d'un ou de plusieurs ganglions lymphatiques) est la forme de tuberculose extrapulmonaire la plus couramment rencontrée. L'adénopathie cervicale (gonflement de ganglions du cou) est la plus commune, mais des atteintes inguinales, axillaires (à l'aisselle), mésentériques (au niveau de l'intestin), médiastinales (entre les poumons et le coeur) et intramammaires ont également été décrites.

La tuberculose pleurale (ou pleurésie tuberculeuse) survient chez jusqu'à 30 % des patients atteints de tuberculose [84]. Cela correspond à un épanchement pleural liquidien, qui est la présence de liquide dans la cavité pleurale, c'est-à-dire le feuillet viscéral adhérant au poumon et le feuillet pariétal adhérant à la cage thoracique.

La tuberculose osseuse (ou maladie de Pott) peut représenter jusqu'à 35 % des cas de tuberculose extrapulmonaire. La tuberculose osseuse implique le plus souvent la colonne vertébrale, suivie par l'arthrite tuberculeuse dans les articulations portantes et l'ostéomyélite tuberculeuse extraspinale (en dehors de la colonne vertébrale).

La tuberculose du système nerveux central comprend la méningite tuberculeuse (la présentation la plus courante), les tuberculomes intracrâniens et l'arachnoïdite tuberculeuse spinale (inflammation chronique d'une fine enveloppe qui entoure le cerveau). Cette forme se distingue par une période initiale caractérisée par un sentiment général de malaise, de céphalées, de fièvre ou d'altérations de la personnalité. Elle est généralement suivie, après deux à trois semaines, par des céphalées persistantes, des symptômes de méningisme, des nausées entraînant des vomissements, une confusion mentale. Sans traitement approprié, le déclin de l'état mental peut mener au coma. Des convulsions peuvent avoir lieu à tous les stades de la maladie [99].

La tuberculose abdominale peut toucher le tractus gastro-intestinal, le péritoine, les ganglions lymphatiques mésentériques. D'autres organes (par exemple, le foie, la rate, les glandes surrénales) sont généralement affectés à la suite d'une tuberculose miliaire.

Enfin, la tuberculose miliaire fait référence à toute forme progressive et disséminée de tuberculose. Elle résulte d'une dissémination par le sang des bacilles de *Mycobacterium tuberculosis* et est caractérisée par de petits tubercules ressemblants en taille et en apparence à des graines de millet. La pandémie mondiale de VIH et l'utilisation généralisée de médicaments immunosuppresseurs ont modifié l'épidémiologie de la tuberculose miliaire. Considérée principalement comme une maladie des nourrissons et des enfants à l'ère préantibiotique, la tuberculose miliaire est de plus en plus rencontrée également chez les adultes [242]. La maladie peut survenir lors de la dissémination d'une infection primaire ou après des années de tuberculose non traitée. La tuberculose miliaire est observée chez 10 % des patients ayant le SIDA et une tuberculose pulmonaire, et chez 38 % de ceux qui ont le SIDA et une tuberculose extra-pulmonaire.

2.4.4 Vaccin BCG

Le seul vaccin homologué actuellement disponible contre la tuberculose est le vaccin Bacille Calmette-Guérin (BCG). Un vaccin, M72/AS01E, actuellement en phase II, semble prometteur. Un second, MTBVAC, est actuellement en phrase III [1]. La vaccination des enfants avec le BCG peut conférer une protection, notamment contre les formes graves de tuberculose chez les enfants [307].

Histoire Dans les années 1900, Albert Calmette et Camille Guérin entament des recherches sur la tuberculose à l'institut Pasteur de Lille. Leur travail incluait la culture de souches virulentes du bacille de la tuberculose et le test de différents milieux de culture. En 1902, Edmond Nocard leur fournit une culture de *Mycobacterium bovis* isolée d'une vache tuberculeuse. Calmette et Guérin remarquent qu'un mélange de glycérine-bile-pomme de terre permettait de faire croître des bacilles qui semblaient moins virulents. Il ont alors essayé de déterminer si la culture répétée pourrait produire une souche suffisamment atténuée pour être envisagée comme vaccin. En 1919, après environ 230 cultures réalisées au cours des 11 années précédentes, ils avaient un bacille tuberculeux qui ne produisait pas de tuberculose lorsqu'il était injecté dans des cochons d'Inde, des lapins, du bétail ou des chevaux. Au départ nommé Bacille Bilié Calmette-Guérin, ils ont omis "Bilié" pour laisser Bacille Calmette-Guérin.

En 1921, Calmette décida que le moment était venu pour un essai du vaccin sur l'homme. La première administration humaine du BCG a été réalisée à l'Hôpital de la Charité, à Paris. Une femme était décédée de la tuberculose quelques heures après avoir donné naissance à un nourrisson en bonne santé. Le 18 juillet 1921, il a été administré une dose de BCG par voie orale au nourrisson. Il n'y a pas eu d'effet indésirable. La voie orale a été choisie puisque Calmette considérait le tube digestif comme la voie habituelle d'infection par le bacille tuberculeux. À partir de 1924 plusieurs campagnes de vaccination furent lancées sur des enfants, sans complications sérieuses [170].

Efficacité Bien que le BCG ait démontré une efficacité significative dans plusieurs populations, la protection n'a pas été constante contre toutes les formes de tuberculose et dans tous les groupes d'âge. En effet, il est montré que le BCG offre une protection constante et appréciable contre la méningite tuberculeuse et la maladie miliaire. Cependant, l'efficacité est variable sur la tuberculose pulmonaire chez les adultes, et il n'existe aucune preuve de son efficacité lorsque le BCG est utilisé comme prophylaxie post-exposition [300, 204].

Il existe plusieurs pistes concernant les raisons de l'inefficacité. Cela peut être expliqué par des différences entre les vaccins BCG. Il est en effet reconnu que les souches produites par différents fabricants diffèrent en propriétés génétiques. D'autre part, la plupart des populations dans le monde sont exposées à diverses mycobactéries "environnementales", et cette exposition peut offrir un certain degré de protection contre une infection ultérieure par les bacilles de la tuberculose que le BCG ne pourrait pas grandement améliorer [170].

Le vaccin n'est pas administrable à toutes les populations. Une administration à un individu immunodéprimé, tel qu'un nourrisson souffrant de déficience immunitaire grave, ou une infection par le VIH, peut entraîner une Bécégite disséminée. De manière générale, la Bécégite disséminée peut survenir entre 1,56 et 4,29 cas par million de doses et présente un taux de létalité élevé [300].

Utilisation actuelle La vaccination par le BCG est recommandée dans les pays ou régions où la tuberculose est très répandue ou où la lèpre est un problème majeur. On peut également envisager la vaccination par le BCG dans les zones où sévit l'ulcère de Buruli (infection nécrosante de la peau). Ces vaccins sont considérés comme efficaces, surtout pour prévenir la méningite tuberculeuse chez les enfants et la tuberculose miliaire. Dans les pays où l'incidence de la tuberculose est faible, il reste possible de vacciner les nouveau-nés dans les groupes de population à haut risque pour la tuberculose. La mise au point de nouveaux vaccins représente une priorité capitale. Il est impératif de concevoir des vaccins offrant une protection supérieure à celle procurée par le BCG, capables de prévenir efficacement toutes les variantes de la tuberculose, y compris celles résistantes aux traitements, ainsi que d'empêcher la réactivation de la maladie. Ces vaccins devraient également se montrer efficaces chez tous les individus, indépendamment de leur âge ou de leur statut sérologique vis-à-vis du VIH, et garantir une performance homogène au sein de diverses populations [300].

2.4.5 Traitements

Il y a trois objectifs dans le traitement de la tuberculose :

— Éliminer rapidement la tuberculose pour produire une amélioration rapide de l'état de santé, réduire la mortalité et éviter une transmission supplémentaire.
— Prévenir l'émergence ou l'aggravation de la résistance aux médicaments.
— Stériliser les lésions pour prévenir les rechutes et permettre une guérison durable.

[217, 137]

Le traitement médicamenteux pour la tuberculose se déroule en deux phases : la phase intensive et la phase de consolidation. Durant la phase intensive, 3 ou 4 médicaments efficaces sont utilisés en combinaison pour tuer rapidement les bacilles de la tuberculose et prévenir la sélection de bacilles résistants aux médicaments. La phase intensive dure 2 mois et pendant cette période les médicaments sont administrés quotidiennement. Dans la phase de consolidation, un minimum de 2 médicaments efficaces sont utilisés, et la durée du traitement varie en fonction du régime médicamenteux, de l'adhérence (participation du patient) et du risque de rechute [137].

Les patients atteints de la tuberculose présentent en général un très grand nombre de bacilles (jusqu'à 10^{10} dans le cas de la tuberculose pulmonaire). À cette concentration, des mutations se produisent spontanément et à un taux faible mais constant. Ces mutations peuvent entraîner des résistances à certains médicaments. Ainsi une résistance spontanée à un médicament est très probable, mais moins probable à plusieurs médicaments.

L'isoniazide (INH), la rifampine (RMP), le pyrazinamide (PZA) et l'éthambutol (EMB) sont des antibiotiques couramment utilisés en première ligne. L'isoniazide est un médicament clé avec une forte activité bactéricide initiale, causant fréquemment une augmentation asymptomatique des enzymes hépatiques, une hépatite clinique, une neuropathie (douleurs chroniques dues à une lésion nerveuse) périphérique, parmi d'autres effets secondaires. La rifampicine est un médicament essentiel, connu pour prévenir la résistance aux médicaments et la rechute, mais nécessite une gestion minutieuse en raison de ses interactions médicamenteuses et d'effets secondaires potentiels comme l'hépatotoxicité. La pyrazinamide (PZA) améliore les effets de stérilisation dans la phase de traitement initiale mais est liée à l'hépatotoxicité et à d'autres effets indésirables, sans affecter les taux de rechute lorsqu'elle est utilisée dans les phases ultérieures. L'éthambutol (EMB) est utilisé principalement pour prévenir la résistance aux médicaments dans le traitement de la tuberculose, le principal problème étant de causer une neuropathie optique parmi d'autres effets secondaires moins courants [137].

La streptomycine (SM) a été le premier médicament utilisé pour le traitement de la tuberculose en 1948. La streptomycine tue les bacilles tuberculeux en croissance active, mais elle est inactive contre les bacilles non croissants ou intracellulaires. L'ototoxicité et la néphrotoxicité sont associées à l'administration de streptomycine.

La tuberculose résistante aux médicaments continue de constituer une menace pour la santé publique (figure 2.19). La résistance à la rifampicine, le médicament de première ligne le plus efficace, est particulièrement préoccupante. La tuberculose résistante à la rifampicine et à l'isoniazide est définie comme une tuberculose multirésistante (MDR-TB). Tant la MDR-TB que la tuberculose résistante à la rifampicine (RR-TB) nécessitent un traitement avec des médicaments de deuxième ligne [205]. Différentes catégories de résistances ont été déterminées en fonction de la résistance à un ou plusieurs médicaments (tableau 2.2). Les dernières directives de l'OMS donnent la priorité à un nouveau régime de 6 mois composé de bédaquiline (B), prétomanide (Pa), linézolide (L) et moxifloxacine (M), appelé BPaLM. Pour les personnes atteintes de tuberculose pré-résistante aux médicaments de deuxième ligne (pré-XDR-TB, définie comme une tuberculose résistante à la R et à toute fluoroquinolone), le régime peut être utilisé sans moxifloxacine (BPaL) [205].

Résistance	Catégorie
Pre-XDR-TB et un groupe A [(a)] hors fluoroquinolone [(b)]	XDR-TB
MDR-TB ou RR-TB et au moins un fluoroquinolone	Pre-XDR-TB
isoniazide, rifampicine	MDR-TB
isoniazide	Hr-TB
rifampicine	RR-TB

TABLE 2.2 – Catégorie de résistance des souches selon les marqueurs de résistance [303, 301].
(a) lévofloxacine, moxifloxacin, bédaquiline, linézolide (b) lévofloxacine, moxifloxacin

2.4.6 Épidémiologie

Bien que la tuberculose est parfois perçue comme une maladie ancienne, ce n'est pas le cas. En 2022, on estime que la tuberculose a causé 1.3 millions de morts dans le monde. On estime par ailleurs que la maladie s'est développée chez 10.6 millions de personnes.

Géographie En 2022, la majorité des personnes ayant développé la tuberculose se trouvaient dans les régions de l'Asie du Sud-Est (46 %), de l'Afrique (23 %) et du Pacifique occidental (18 %), avec des proportions plus faibles dans la région de la Méditerranée orientale (8,1 %), des Amériques (3,1 %) et de l'Europe (2,2 %). Trente pays représentent 87 % des cas de tuberculose dans le monde et les deux tiers du total mondial sont dans huit pays : l'Inde (27 %), l'Indonésie (10 %), la Chine (7,1 %), les Philippines (7,0 %), le Pakistan (5,7 %), le Nigeria (4,5 %), le Bangladesh (3,6 %) et la République démocratique du Congo (3,0 %).

La gravité des épidémies nationales de tuberculose, en termes de nombre de cas incidents de tuberculose pour 100 000 habitants par an, varie considérablement d'un pays à l'autre, allant de moins de 10 à plus de 500 nouveaux cas et cas de rechute pour 100 000 habitants par an (figure 2.18).

Répartition sociale La répartition en termes de sexe est inégale, en 2022, 55 % des personnes ayant développé la tuberculose étaient des hommes et 33 % étaient des femmes, et 12 % sont des enfants de moins de 14 ans. Environ 50 % des patients atteints de tuberculose font face à des coûts totaux (dépenses médicales directes, dépenses non médicales et coûts indirects tels que les pertes de revenus) qui sont supérieurs à 20 % du revenu annuel du ménage. Parmi tous les cas de tuberculose en 2022, 6,3 % étaient des personnes vivant avec le VIH. La proportion de personnes ayant le VIH et étant nouvellement touché par la tuberculose était la plus élevée dans les pays d'Afrique, dépassant 50 % dans certaines parties de l'Afrique australe.

2.5 Complexe *Mycobacterium tuberculosis* (MTBC)

La tuberculose chez les humains et chez les animaux est causée par un groupe de mycobactéries appelé *Mycobacterium tuberculosis* complex (MTBC). Les différentes lignées du MTBC sont similaires à 99.9% au niveau nucléotidique, avec une distance génétique maximale entre deux souches d'environ 2500 SNP [48].

Le MTBC comprend 7 lignées pour lesquelles l'humain est le seul hôte connu. Ces lignées comprennent *Mycobacterium tuberculosis sensu strictu* (lignées 1-4 et 7), ainsi que *Mycobacterium africanum* (lignées 5 et 6). La distribution géographique de ses lignées diffère, avec certaines lignées présentes globalement dans le monde et d'autres seulement dans certaines régions géographiques.

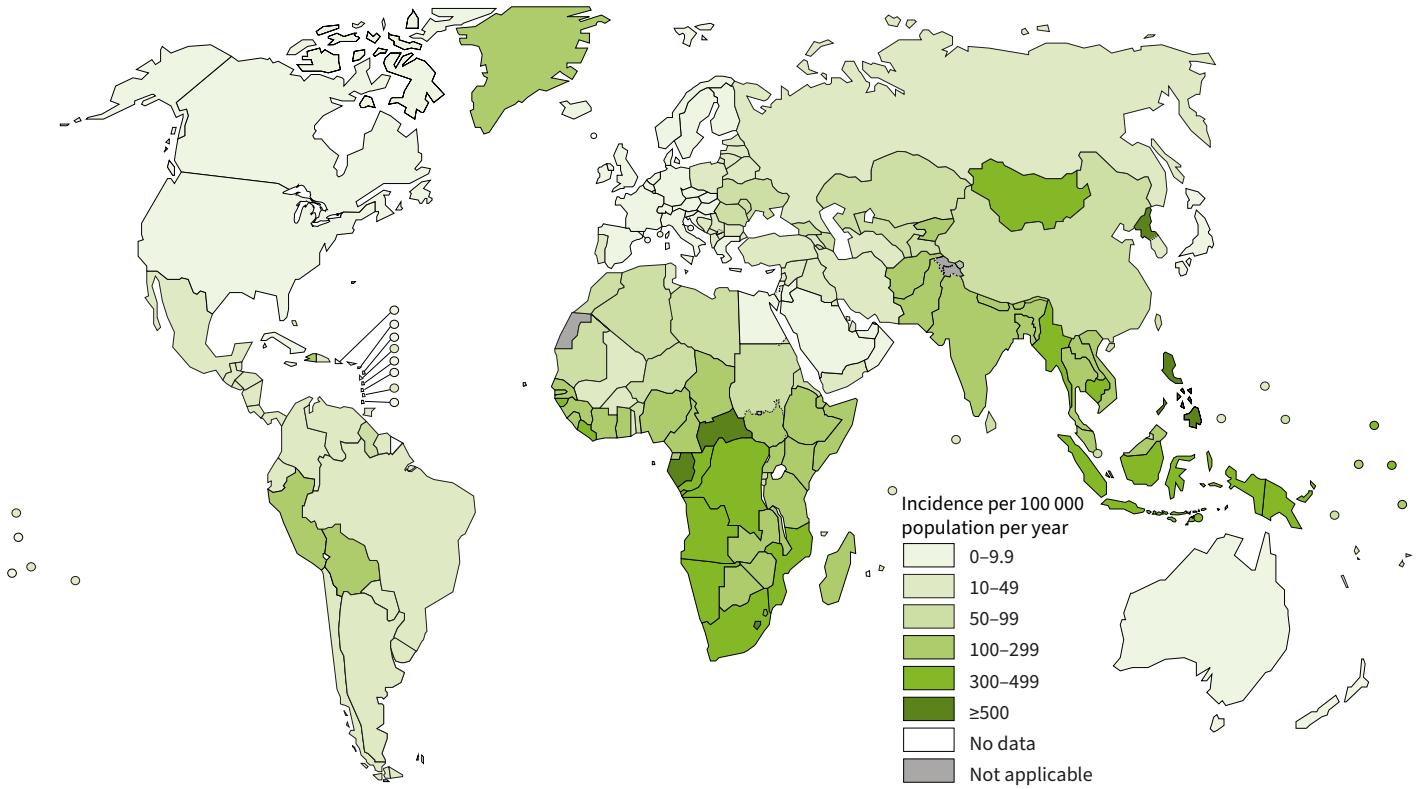


FIGURE 2.18 – Taux d'incidence estimé de la tuberculose en 2022 [205]

FIGURE 2.19 – Nombre estimé de personnes ayant développé une tuberculose multirésistante ou résistante à la rifampicine (cas incidents) en 2022, pour les pays comptant au moins 1000 cas incidents. Les huit pays classés par ordre décroissant du nombre total de cas incidents de tuberculose résistante aux médicaments (RR-TB) en 2022 sont l'Inde, les Philippines, la Fédération de Russie, l'Indonésie, la Chine, le Pakistan, le Myanmar et le Nigeria [205].

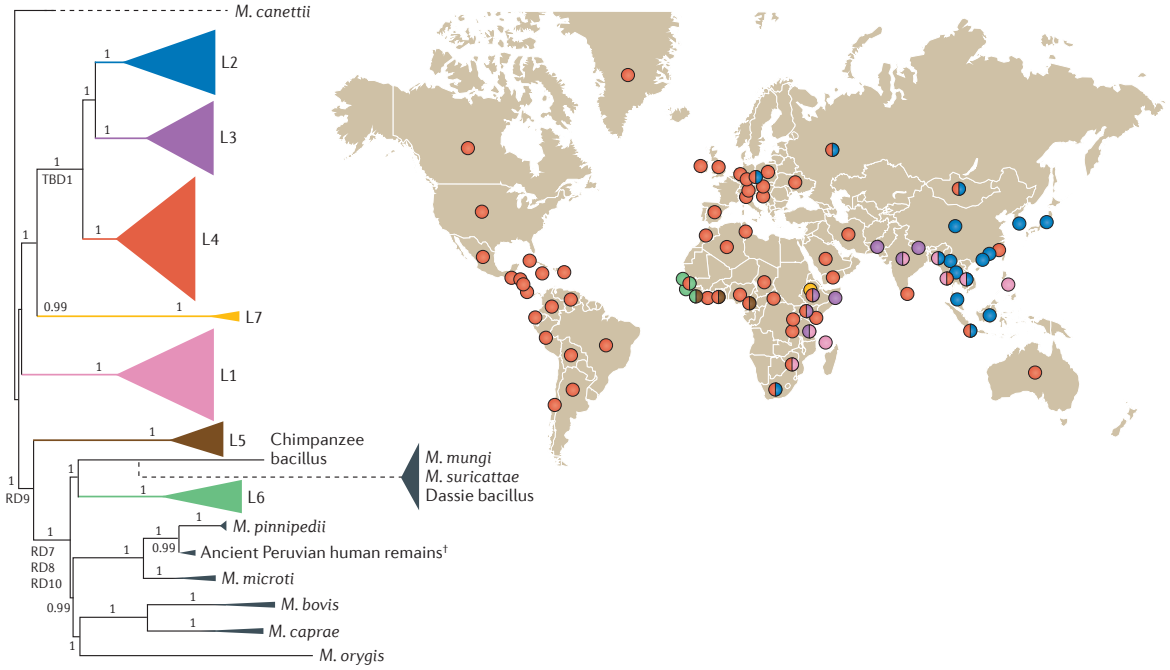


FIGURE 2.20 – Arbre phylogénétique du MTBC et répartition géograpique [92]

2.5.1 *Mycobacterium tuberculosis sensu strictu*

Mycobacterium tuberculosis sensu strictu comprend au total 5 lignées.

— Les lignées 2 (lignée d'Asie de l'Est) et 4 (lignée Euro-Americaine) sont présentes dans le monde entier.

— La lignée 3 (Souche d'Asie centrale ou CAS) est dominante dans des parties de l'Afrique de l'Est, en Asie Centrale et au nord de l'Asie du Sud.

— La lignée 1 (lignée Indo-Océanique ou EAI) est principalement présente autour de l'océan indien et en Afrique de l'Est (principalement au Mozambique et à Madagascar).

— La lignée 7 est originaire d'Éthiopie, où elle a été identifiée originalement dans la région Woldiya [201].

Les lignées 2, 3 et 4 sont qualifiées de modernes et les autres sont qualifiées d'ancestrales [24]. Malgré leur connotation temporelle, ces termes font référence à l'absence ou à la présence de la région TbD1 dans leur génome, et à une classification utilisée en 1997 par Sreevatsan et al. [257]. Les lignées modernes sont responsables des épidémies globales alors que les autres ont tendance à être restreintes à des régions spécifiques [21].

2.5.2 *Mycobacterium africanum*

Mycobacterium africanum est principalement présente en Afrique de l'Ouest où elle cause la moitié des morts de la tuberculose.

— On y distingue une première lignée, *West African 1* ou MAF1 (lignée 5) principalement présente autour du Golfe de Guinée

— Et une seconde, *West African 2* ou MAF2 (lignée 6) principalement présente à l'Ouest de l'Afrique de l'Ouest.

— Plus récemment, la lignée 9 a été trouvée uniquement à Djibouti et en Somalie [61].

Auparavant il y avait deux lignées *M. africanum type I* et *II*. *M. africanum type II* a été reclassée récemment en *M. tuberculosis sensu stricto* (une sous-division appelée Uganda) par les avancées dans les techniques de génotypage. Et MAF1 et MAF2 sont des sous-divisions de *M. africanum type I* [70].

2.5.3 Autres lignées

Le MTBC comprend aussi une lignée plus distante des autres, connue comme *M. canettii*, et autres smooth tuberculosis bacilli (STB). Ils sont caractérisés par leur morphologie de colonie lisse [57].

Récemment, deux nouvelles lignées ont été proposées. La lignée 8, rare et réduite aux Grands Lacs d'Afrique [202] et la lignée 9 déterminée par des différences importantes avec les autres lignées à partir de souches prélevées sur des patients en Afrique [59]. Enfin, la lignée 10, dont la découverte est issue du travail décrit dans ce manuscrit [110].

On distingue aussi d'autres lignées adaptées à différentes espèces comme *M. microti* (mulot sylvestre, campagnols), *M. pinnipedii* (mammifères marins), *M. caprae* (associé aux chèvres et cerfs), *M. bovis* (bétail), *M. orygis* (Oryx) [249], *M. suricattae* (suricate) [210] ou encore le chimpanzé [58].

Une partie de ces lignées est, dans de rares cas, la cause de tuberculose chez l'humain. Par exemple *M. orygis* [286], *M. bovis* via du lait contaminé ou un contact [195], *M. caprae* [195], et *M. microti* dans de très rares cas [123].

2.6 Marqueurs génétiques et génotypage du MTBC

2.6.1 Séquences d'insertion

Les souches du MTBC comptent de nombreuses séquences d'insertion. Par exemple, la souche de référence H_{37}Rv contient 16 copies de la séquence d'insertion IS*6110*, 6 copies de IS*1081* ainsi que 32 autres IS distinctes [41].

L'IS*6110* est particulièrement intéressante car il a été découvert qu'elle est spécifique au MTBC. Elle permet ainsi d'être utilisée dans la détection du MTBC via PCR pour un diagnostic rapide [278]. Cependant, il a été découvert par la suite que certaines souches, démontrées comme étant du MTBC par séquençage de la séquence ARNr 16S ou d'autres méthodes, ne possèdent pas cette IS [288, 169]. On fait souvent référence aux séquences IS*987* et IS*986* par la désignation IS*6110* car ces séquences sont identiques à quelques nucléotides près [284]. On note aussi que l'IS*6110* est presque systématiquement présente dans le locus CRISPR [102].

En général, le nombre de copies d'IS*6110* varie de 0 à 25 et dépend de la fréquence de la transposition, qui est largement conditionnée par la nature de la région génomique où se produit la transposition [134]. Bien que l'IS*6110* puisse s'insérer à n'importe quel point du génome, il existe des régions où la fréquence de transposition est plus élevée. Les zones d'intégration privilégiées, appelées "hot spots", sont généralement situées dans les régions codantes. La variation en nombre et en position dans le génome de l'IS*6110*, ainsi que le fait que leur transposition soit un événement rare [287], permet la discrimination de différentes souches du MTBC, ce qui a conduit la mise au point d'une méthode de fingerprinting polymorphisme de longueur des fragments de restriction (RFLP) [284]. Cette méthode pose cependant problème dans le cas de populations ayant un faible nombre de copies de l'IS*6110*, comme c'est le cas dans certaines parties de l'Asie comme en Inde [69]. Dans ce cas une méthode de typage secondaire est parfois utilisée, comme le polymorphisme des séquences répétitives riches en GC (PGRS)[226] appelé PGRS-RFLP. Le typage RFLP est techniquement exigeant et requiert une quantité importante d'ADN, nécessitant la culture de la bactérie. Cela est à comparer avec le spoligotyping (section 2.6.3) qui présente de nombreux avantages, même s'il a été constaté que le niveau de différenciation apporté par le spoligotyping était parfois plus faible que pour la méthode RFLP [176].

2.6.2 Régions de différence

Les analyses comparatives du génome ont montré de large variations structurales parmi les souches MTBC. Ces régions sont couramment appelées regions of difference (RDs). La

suppression de ces régions se comporte parfois comme un unique event polymorphism (UEP) [120], un événement de mutation unique et donc probablement conservé parmi les souches. Ces RDs ont par conséquent été utilisées comme marqueurs phylogénétiques [24].

— La nomenclature a été initiée par Mahairas et al. [172] avec le travail sur le vaccin BCG, dérivé de *M. bovis*. Cela a conduit à la découverte de trois régions supprimées des souches BCG (RD1-RD3) mais présentes dans *M. tuberculosis* et *M.bovis*. RD1 fait partie du locus ESX-1 qui code un système de sécrétion de type VII, impliqué dans la virulence et la pathogenèse de la tuberculose [211]. La suppression de RD2 s'est produite entre 1927 et 1931 à l'Institut Pasteur [16], et est connue pour conduire à une immunogénicité accrue [151].

— RD4 découvert par Brosch et al. [25] est absent dans les souches communes de *M.bovis* et de BCG et joue un rôle dans la virulence mycobactérienne [228].

— RD5-RD10 correspondent à la nomenclature de Gordon et al. [100]. RD5 et RD6 sont des séquences d'insertion IS*6110* et IS*1532* respectivement. RD7-RD10 sont tous présents dans *M. tuberculosis*. RD9 est supprimé dans *M. africanum* et les lignées animales. RD7, RD8, RD10 sont supprimées des souches animales et de la lignée 6 [24].

— RD11-14 ont été décrites par Behr [16]. RD11 correspond à un élément génétique mobile, et RD12-14 sont reconnus comme marqueurs phylogénétiques pour les sous-lignées de *M. bovis* [24].

Un certain nombre d'études ont complété cette liste et identifié plus de 200 Régions de Différence (tableau B.1), certaines d'entre elles se sont avérées être des marqueurs génétiques fiables pour identifier précisément les sous-lignées MTBC. Constituer une liste précise de toutes les RD implique d'adapter les positions décrites dans les différents articles à la souche de référence utilisée. Dans certains cas (par exemple Mahairas et al. [172]) une jonction PCR est donnée afin de situer précisément la zone supprimée du génome. Deux amorces de chaque côté de la zone supprimée sont utilisées afin d'amplifier la séquence située entre les deux amorces. Après séquençage nous pouvons constater ou non la présence de la délétion. Et nous pouvons utiliser ces séquences jonction afin de trouver les positions précises des délétions dans des souches de référence (figure 2.21).

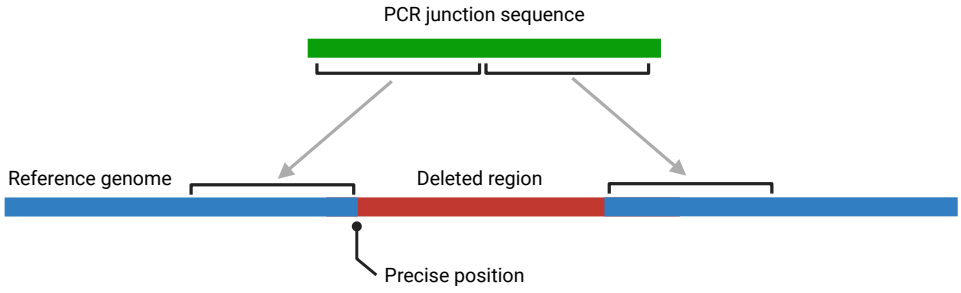


FIGURE 2.21 – Utilisation de la séquence de jonction pour déterminer les positions exactes des RD dans le génome

2.6.3 CRISPR

En 1987, des séquences répétitives non formalisées à cette date ont été découvertes dans la souche K12 de *Escherichia coli* [198], et plus tard dans d'autres bactéries. C'est seulement récemment que ces séquences furent reconnues comme une famille à part entière [186]. Les auteurs nommant cette famille sous des noms différents, et entraînant une nomenclature confuse, il a été ensuite décidé de nommer cette famille clustered regularly interspaced short palindromic repeats (CRISPR)[135]. Quelques années plus tard, des analyses *in silico* ont révélé que l'on trouve un CRISPR dans près de 40% des génomes de bactéries séquencés. On retrouve des gènes souvent associés au CRISPR, dits

CRISPR-associated (CAS).

D'un point de vue moléculaire, les CRISPR sont des répétitions directes dispersées, mais leur particularité est que les répétitions sont espacées par des séquences non répétitives (appelées *spacers*) qui ont une taille similaire aux répétitions. La taille des répétitions directes observées peut varier de 21pb (Salmonella typhimurium) à 37pb (Streptococcus pyogenes), et les CRISPR peuvent se trouver à un ou plusieurs loci dans un génome.

En 2007, Barrangou et al. [15] ont démontré expérimentalement que le CRISPR agissait comme système de défense anti-phages. Les bactériophages, abrégé phages, étant des virus qui n'infectent que des bactéries. En réponse à l'infection d'un phage, la bactérie intègre de nouveaux *spacers* dérivés de la séquence du génome du phage. Il a été découvert que lorsqu'un spacer avait une séquence identique à celle d'un phage, la bactérie devenait résistante à ce phage. Le système CRISPR-Cas est donc adaptatif dans le sens où il a la capacité d'acquérir une nouvelle immunité suite à l'infection. Cette adaptation est permise par les protéines Cas1 et Cas2. [256, 9, 173]

Les souches du MTBC contiennent une région chromosomique unique CRISPR initialement identifiée par Hermans et al. [114] dans la souche vaccinale *M. bovis* BCG P3. Elle est caractérisée par de multiples direct repeat (DR) de 36pb espacées par des séquences uniques de 35pb à 41pb. Il a été constaté que les souches du MTBC n'acquièrent plus de nouveaux *spacers*, ce qui est en adéquation avec le fait que Cas1 et Cas2 n'ont pas été détectés dans le protéome de *M. tuberculosis*. Aussi, il n'est pas démontré que *M. tuberculosis* acquiert une résistance aux phages via le système CRISPR [314].

DVR-PCR

En 1993, la région CRISPR du MTBC, alors appelée locus direct repeat (DR), a été utilisée pour la première fois pour la différentiation des souches[102] de par son polymorphisme, lui déjà connu depuis 1991[114]. La méthode appelée direct variable repeat polymer chain reaction (DVR-PCR), dérivée de la méthode minisatellite variant repeat polymer chain reaction (MVR-PCR) [136], permet le typage de souches en une seule PCR.

Cependant cette méthode n'était pas adaptée à une utilisation dans un laboratoire de biologie médicale de par sa difficulté technique.

Spoligotyping

En 1997, est présentée une nouvelle méthode de PCR et d'hybridation appelée spoligotyping (pour spacer oligotyping)[138]. Cette méthode permet de détecter la présence ou l'absence d'un ensemble de 43 *spacers*.

Dans une première phase d'amplification par PCR (figure 2.23), deux oligonucléotides DRa et Drb (tableau 2.3), basés sur le DR constant (figure 2.22), sont utilisés comme amorces. L'amorce DRa est biotinylée pour permettre ensuite une visualisation immunohistochimique.

Amorce	Séquence
DRa	GGTTTTGGGTCTGACGAC (5' biotinylée)
DRb	CCGAGAGGGGACGGAAAC

TABLE 2.3 – Séquences des amorces DRa et DRb utilisées pour le spoligotyping

Dans une seconde phase, le produit de la PCR est ensuite hybridé avec 43 oligonucléotides correspondant à 43 *spacers* uniques du CRISPR. Ces 43 *spacers* ont été déterminés à partir de la souche $H_{37}Rv$ (*M. tuberculosis*) et BCG P3 (*M. bovis*).

Pour cela, les oligonucléotides des 43 *spacers* ont été liés de manière covalente et linéairement à une membrane. Un MiniBlotter est ensuite placé perpendiculairement aux lignes de la membrane. Chaque sillon est ensuite rempli avec le produit PCR, préalable-

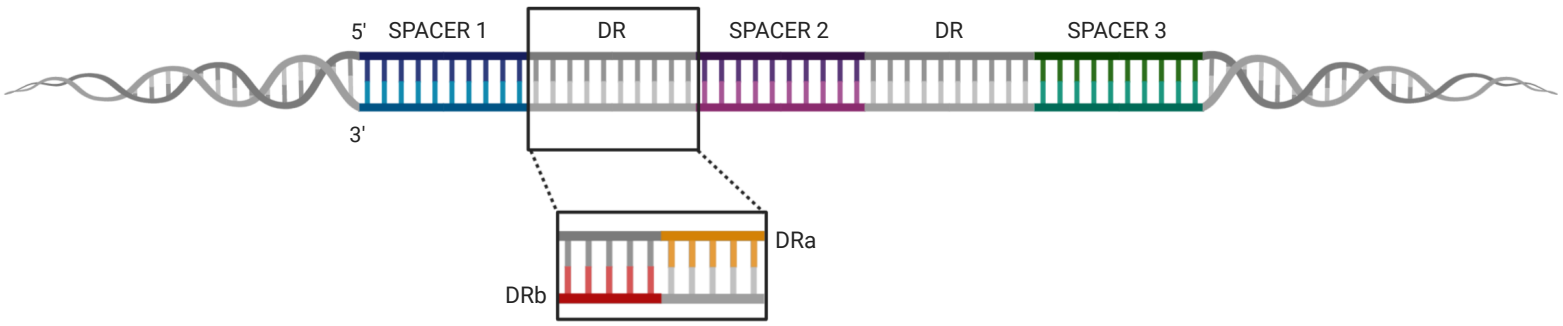


FIGURE 2.22 – Séquence du CRISPR avec DRa et DRb, les oligonucléotides qui serviront d'amorces pour la PCR

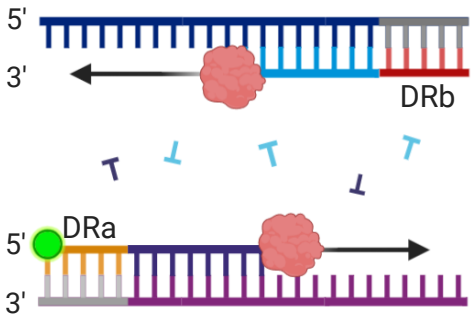


FIGURE 2.23 – PCR avec les amorces DRa (biotinylée) et DRb

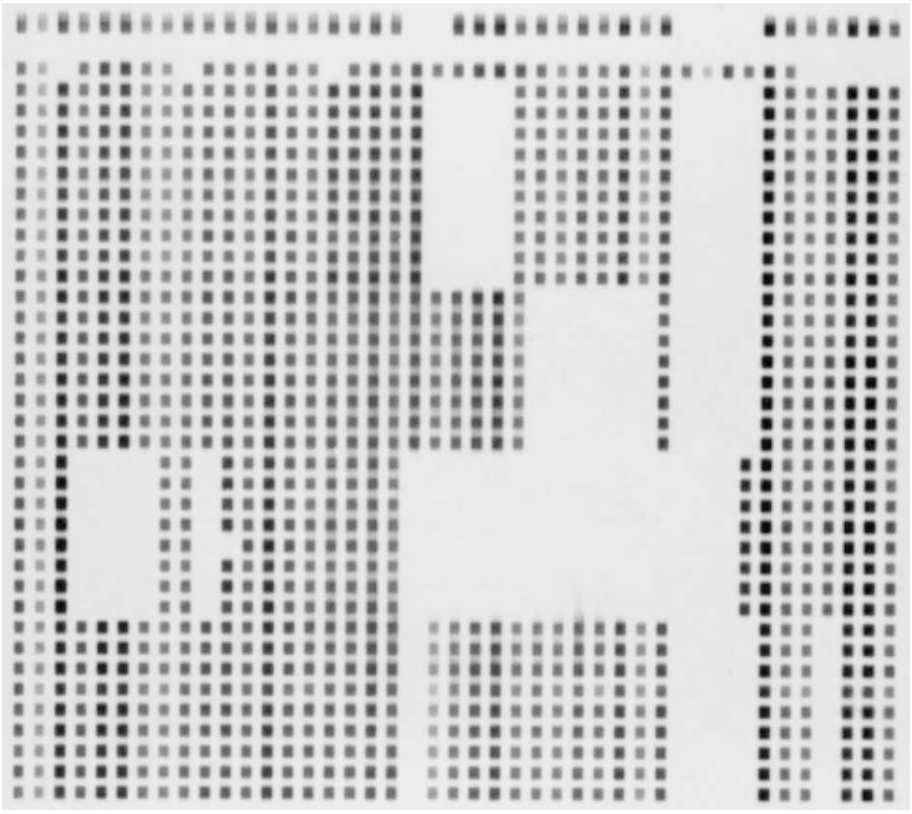
ment dénaturé, d'une souche (figure 2.24b). Cela permet de réaliser le spoligotyping de 45 échantillons en même temps dans le cas d'un MiniBlotter 45 [128].

Il s'ensuit l'hybridation du produit de la PCR avec les oligonucléotides de la membrane. Si un spacer est présent dans le produit de la PCR, il viendra se lier à la ligne correspondante de la membrane.

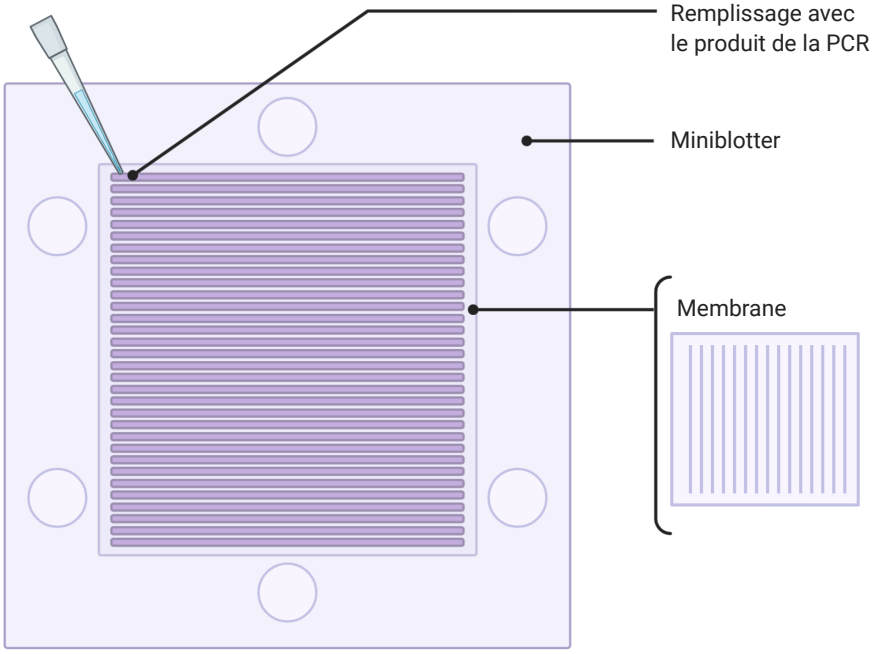
En utilisant le marquage à la biotine, on peut ensuite relever la présence des *spacers* à l'intersection des lignes par chimiluminescence, qui est la production de lumière à la suite d'une réaction chimique (figure 2.24a).

De par sa forme, le résultat du spolygotyping (appelé spoligotype) est facilement représentable. Sous forme binaire par exemple, avec une suite de 43 bits, 1 codant la présence d'un spacer et 0 son absence. Ainsi en 1999, est publiée une première base de données «SPOLDB1» regroupant 610 spoligotypes [251], suivie de plusieurs autres bases de données dont «SITVITWEB»[72] et «SITVIT2»[63] comprenant les spoligotypes de plus de 110 000 isolats.

Il a été démontré que les *spacers* utilisés dans le spolygotyping présentent de l'homoplasie (des mutations indépendantes qui résultent en la perte du même spacer), ce qui rend le spolygotyping un outil potentiellement moins fiable pour les analyses phylogénétiques formelles. L'homoplasie dans les *spacers* pourrait conduire les spoligotypes de souches non apparentées à converger vers des motifs identiques [145]. D'autre part, le spolygotyping ne permet pas une résolution au niveau de la souche, mais fournit plutôt des informations au niveau de la sous-lignée ou de la lignée. Pour exemple, la plupart des souches de la lignée de Beijing, une lignée importante avec une large distribution géographique, partagent un unique spoligotype composé des neuf derniers *spacers* [180].



(a) Visualisation du résultat du spoligotyping par ECL sur Hyperfilm ECL[138]



(b) Dépôt du produit de la PCR sur la membrane, perpendiculairement aux oligonucléotides des 43 spacers

FIGURE 2.24 – Visualisation de la méthode de spoligotyping

2.6.4 VNTR

Un variable number tandem repeat (VNTR) est un emplacement dans un génome où une courte séquence de nucléotides est organisée sous forme de répétition en tandem (figure 2.25).

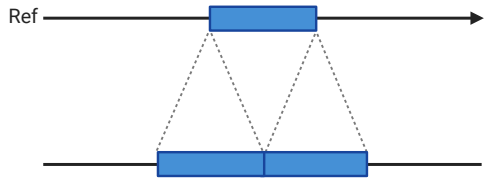


FIGURE 2.25 – Répétition en tandem

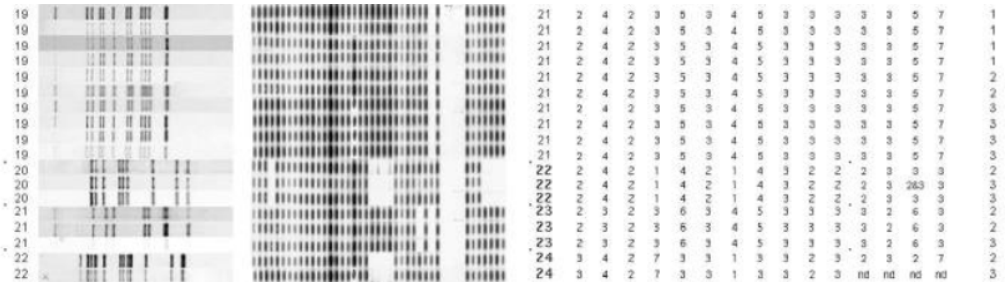


FIGURE 2.26 – IS*6110* RFLP, spoligotype, et MIRU-VNTR d'isolats de *M. tuberculosis* [271]

Méthode de génotypage

Le principe des système de génotypage basés sur les VNTR est basé sur l'amplification des loci VNTR par PCR. La PCR est effectuée en utilisant des amorces complémentaires aux régions entourant les loci VNTR. Le produit de la PCR est ensuite visualisé par électrophorèse sur gel qui permet d'estimer la taille des amplicons. La comparaison de cette taille par rapport à la taille connue d'une répétition permet de déterminer le nombre de répétitions. Le résultat est ensuite exprimé dans un format numérique simple, dans lequel chaque chiffre représente le nombre de copies associé à un locus particulier [134].

MPRT et ETR

Le premier VNTR décrit fut le major polymorphic tandem repeat (MPTR) par Hermans et al. [115]. Le MPTR est composé de répétitions de 15 paires de bases avec une séquence consensus unique, mais avec une variabilité de séquence substantielle entre les répétitions adjacentes. Les répétitions étant légèrement hétérogènes en séquence, c'est pour cela que la séquence à été désignée MPTR. Au total, 5 loci MPTR furent décrits par la suite, nommés MPTR-A à MPTR-E [90].

Les exact tandem repeats (ETRs), présents exclusivement dans les souches du MTBC, contiennent des répétitions dont la taille varie de 53 à 79 bp. Le séquençage des ETRs a montré que tous les loci ETR étaient variables en nombre de répétitions. En comparaison, parmi les loci MPTR, seul un (MPTR-A) a montré un polymorphisme [134].

Sur les 11 loci MPTR/ETR, seulement cinq (ETR-A à E) sont utilisés pour le génotypage des souches de *M. tuberculosis*. Bien qu'il soit rapide, simple à réaliser et hautement reproductible, le génotypage VNTR basé sur 5 loci ETR présente un pouvoir discriminatoire faible, comparé à l'IS*6110*-RFLP ou au spoligotypage.

MIRU

Une autre classe de VNTR sont les mycobacterial interspersed repetitive units (MIRUs), décrits pour la première fois par Supply et al. [268]. Il s'agit de 41 loci de répétitions en tandem d'une taille comprise en 46 et 101pb. Sur ces 41 loci, 12 présentent des variations dans le nombre de répétitions et sont ainsi utilisées pour le génotypage de *M. tuberculosis*. Dans ces 12 loci, MIRU-4 et MIRU-31 correspondent aux ETRs ETD-D et ETR-E précédemment décrits. Par la suite, un nouveau système à 24 loci VNTR (incluant les 12 précédemment découverts) fut proposé [271]. L'usage de 4 loci additionnels, appelés «hypervariables», permet en plus de différencier les souches Beijing.

Un code, appelé MIRU-VNTR, associe à chaque locus le nombre de répétitions qui ont été trouvées (figure 2.26). Ce code permet de comparer rapidement les souches et permet d'alimenter des bases de données pour des études épidémiologiques à grande échelle. Parmi les bases de données les plus importantes, on compte «MIRU-VNTRplus» [299] ainsi que «SITVITWEB»[72].

Le pouvoir discriminant du MIRU-VNTR à 24 loci égale celui de l'IS*6110*-RFLP [134]. Même si les loci pris séparément peuvent montrer une homoplasie, l'information phylogénétique transmise par les MIRU-VNTR à 24 loci est moins sujette à l'homoplasie que le spoligotypage [180]. Grâce à sa nomenclature numérique simple, le système MIRU-VNTR basé sur les 24 loci additionnés des 4 loci hypervariables, a remplacé au cours de la dernière décennie la méthode IS*6110*-RFLP [180].

2.6.5 SNP

On distingue deux grandes familles de SNP. Les synonymes, où la mutation de la séquence d'ADN ne modifie pas la séquence d'acides aminés. Et les SNPs non synonymes, où la mutation de la séquence d'ADN entraîne aussi une modification de la séquence d'acides aminés, ce qui peut ensuite entraîner des changements dans les protéines et dans le phénotype de l'organisme. Un SNP non-synonyme peut par exemple altérer la résistance aux antibiotiques du MTBC, et donc une pression de sélection. Par conséquent les SNPs non synonymes ne constituent en général pas les meilleurs marqueurs phylogénétiques [134]. Les SNPs synonymes sont phénotypiquement neutres et ont un faible niveau d'homoplasie [263], ce sont par conséquent des marqueurs robustes dans l'étude des relations phylogénétiques des différentes souches du MTBC.

La région du génome où se trouve le SNP a aussi une importance. Par exemple, dans les régions codantes, les deletions et insertions sont moins probables. Cela peut être expliqué par le fait que les mutations dans ces régions peuvent empêcher la survie de la bactérie [45]. Les mutations sont aussi moins probables dans les gènes essentiels [264]. On préférera donc prendre en marqueur des SNPs se trouvant dans des gènes essentiels ou des régions codantes.

2.7 Whole-genome sequencing (WGS) du MTBC

Les méthodes de génotypage classiques possèdent des limitations importantes. Même le standard du génotypage, MIRU-VNTR à 24 loci, interroge seulement un nombre limité de régions polymorphiques du MTBC. Ces marqueurs ne peuvent suffire à déterminer un profil génomique complet. Par exemple, ces méthodes ne permettent pas de distinguer des souches génétiquement proches et ont tendance à minimiser la distance génétique réelle entre deux souches. D'autre part les méthodes classiques ne permettent pas de capturer les mutations dans des gènes spécifiques, comme ceux associés à la résistance aux médicaments. L'application du WGS pour l'analyse des souches du MTBC offre quant à lui la source d'informations la plus riche sur un isolat clinique donné.

Après séquençage d'un génome, il est possible de le comparer avec une souche de référence connue (en général $H_{37}Rv$) et ainsi obtenir un ensemble de différences génomiques

entre deux souches. Ces variations sont alors utilisées comme marqueurs pour répondre à un grand nombre de questions sur un isolat clinique donné.

Il est par exemple possible de détecter l'intégralité des marqueurs génomiques de résistance aux médicaments. Les cliniciens peuvent obtenir rapidement des informations pertinentes sur le meilleur traitement à adopter, recevant des informations sur la sensibilité aux médicaments de première ligne, ainsi qu'aux médicaments de deuxième ligne et aux nouveaux traitements. À mesure que le nombre de loci identifiés par WGS comme étant liés à la résistance augmente, la valeur de cette approche s'accroit, notamment pour une utilisation en routine de laboratoire. Les recherches récentes ont révélé que plus de 100 régions génétiques jouent un rôle dans la résistance aux médicaments, avec des mutations significatives dans ces zones. Le WGS est donc préconisé pour une analyse détaillée, après validation des mutations par des tests de concentration minimale inhibitrice et d'échange allélique, et en tenant compte de leur corrélation avec les résultats cliniques. Le WGS permet d'identifier rapidement les mutations résistantes connues, aidant ainsi à personnaliser les traitements. Il offre aussi un avantage sur les outils moléculaires recommandés par l'OMS, en fournissant des détails sur les résistances phénotypiques pour des SNPs spécifiques et en analysant des régions génomiques étendues, au-delà des zones habituellement ciblées [38].

Le WGS montre aussi un grand potentiel pour l'étude de la transmission de la tuberculose. Jusqu'à présent, l'utilisation du WGS a principalement été restreinte aux milieux de la recherche et a souvent été mise en œuvre de manière rétrospective. Cette méthode a prouvé sa capacité à déterminer avec une précision accrue les clusters de transmission, dépassant ainsi les limites des techniques de génotypage traditionnelles en évitant les erreurs d'attribution au sein de ces clusters de transmission [49]. Walker et al. [294] ont publié la première étude à grande échelle sur la transmission de la tuberculose basée sur le WGS, démontrant sa supériorité sur les méthodes antérieures de typage des souches en identifiant plus précisément les cas faisant partie d'un foyer épidémique. Ils ont utilisé une distance génétique de cinq SNPs ou moins pour définir des clusters de transmission de haute confiance, souvent corroborés par des liens épidémiologiques, tandis qu'une distance de 5 à 12 SNPs indiquait une transmission récente probable mais moins souvent confirmée épidémiologiquement. Les distances supérieures à 12 SNPs servaient à distinguer les cas épidémiologiquement non liés [49]. Une étude ultérieure menée en Suisse a révélé que le génotypage MIRU surestimait le taux de transmission parmi les immigrants, probablement en raison de la similitude génétique élevée des souches dans les pays à forte prévalence [265].

Une des mesures les plus importantes des essais cliniques de nouveaux médicaments est le nombre de rechutes après un traitement réussi. Dans les pays à forte charge, il est difficile de différencier un véritable cas de rechute, indiquant un échec du médicament/régime, ou une réinfection après la fin du traitement. Les outils permettant de distinguer ces deux cas sont donc essentiels. Plusieurs exemples récents ont mis en évidence le potentiel WGS pour effectuer la distinction entre les rechutes et les infections secondaires. Par exemple, Guerra-Assunção et al. [104] ont utilisé une différence de ≤ 10 SNPs pour définir une rechute, et une différence de > 1000 SNPs a été utilisée pour définir une réinfection. Plus récemment, Liu et al. [164] ont utilisé une différence de 6 SNP pour faire la distinction entre les deux cas.

Enfin, pour définir de manière fiable les associations phylogénétiques entre les souches, les marqueurs génétiques doivent être uniques et, idéalement, irréversibles. De telles mutations phylogénétiquement informatives ont été identifiées chez *M. tuberculosis* sous la forme, par exemple, de SNPs. La fréquence relativement faible de SNP et le transfert horizontal limité en cours dans MTBC entraînent des niveaux faibles d'homoplasie, deux occurrences indépendantes d'une SNP étant peu probable. En utilisant le WGS, il est ainsi possible, après comparaison avec un génome de référence *in silico*, de collecter l'ensemble de ces SNPs présents dans le génome. Et ainsi il est possible de reconstruire une phylogénie

du MTBC [94].

2.7.1 Extraction, préparation et séquençage

L'analyse WGS d'un échantillon d'ADN d'une souche du MTBC implique au moins trois étapes. Dans un premier temps l'ADN est extrait et préparé pour le séquençage. Dans un second temps, la séquence nucléotidique est obtenue via, par exemple, NGS. Enfin, cette séquence est analysée *in silico*.

Au début du NGS, une contrainte majeure pour le WGS des bactéries du MTBC était le temps nécessaire pour obtenir des quantités suffisantes d'ADN pour les protocoles subséquents de préparation de la librairie. Cela nécessitait plusieurs semaines de culture à partir des échantillons cliniques en raison de la croissance lente des organismes. De nouvelles méthodes de préparation ont ensuite été développée, entraînant une réduction considérable du temps nécessaire et des coûts globaux [180]. Il a été montré que l'analyse de routine complète via WGS était possible, et ce avec un temps de réponse plus court et un coût inférieur. Par exemple, Pankhurst et al. [208] ont montré un coût de 481 £ par analyse, incluant le temps du personnel, les consommables et l'équipement, soit 7% de moins d'un diagnostique classique, lui évalué à 518 £.

La technologie NGS (ou de troisième génération) utilisée ensuite déterminera le coût, la longueur des séquences lues ainsi que le taux et le type d'erreur. La majorité des données de séquençage mise en ligne publiquement ces dernières années a été obtenue via la technologie Illumina avec une taille de read comprise entre 50 et 300 paires de bases. Le choix de la technologie dépend aussi du type d'analyse à réaliser. Par exemple, les approches produisant des reads longs comme PacBio seront préférées pour reconstruire un génome complet. Ce genre de méthode est néanmoins plus adapté à des questions de recherche que du typing WGS en raison du coût associé [180].

2.7.2 Stockage des séquences

Le résultat du séquençage est obtenu sous forme de fichiers contenants des fragments de séquences appelés «reads». Les formats et les caractéristiques de ces fichiers, ainsi que les différentes approches pour les étudier, sont très variés.

Le format FASTA (ou format Pearson) est un format de fichier texte utilisé pour stocker les séquences nucléiques. Ces séquences sont représentées par une suite de lettres codant les acides nucléiques : T, A, G et C. Chaque séquence correspond à un read, et chaque séquence peut être précédée par un nom et des commentaires. Le format FASTA est un des formats les plus simple et lisible, il est supporté par la majorité des outils d'analyse. Les séquenceur produisent en général des fichiers FASTQ, qui est une extension du format FASTA. FASTQ associe à chaque base un score de qualité, appelé «phred», qui est déterminé par le séquenceur en fonction de la probabilité qu'il s'agisse d'une mauvaise identification de la base nucléique. Les fichiers FASTA ou FASTQ sont en général compressés, car ils peuvent contenir des millions de séquence et leur taille peut être de l'ordre de plusieurs gigaoctets.

Les fichiers FASTA sont bien souvent par paires dans le cas du séquençage paired-end. Le fait de lire les deux extrémités des segments d'ADN permet un alignement plus précis, c'est pour cela que le paired-end est souvent privilégié pour des analyses WGS nécessitant une plus grande précision.

De manière générale, les fichiers de séquence peuvent être publiés en accès libre sur des plateformes, comme celle du National Center for Biotechnology Information (NCBI), qui contient des millions de séquences publiques.

2.7.3 Contrôle de la qualité

Filtrage

Une des principales étapes pour s'assurer de la qualité des données de séquençage est le filtrage des reads suivant plusieurs critères. Cette étape vise à éliminer les reads de faible qualité et pouvant impacter les analyses des données.

Un premier critère de filtrage est le score phred. Les reads sont filtrés en fonction de leur pourcentage de base disposant d'un score phred faible. Il est aussi possible de filtrer le read en fonction du score phred moyen. Un second critère de filtrage est la taille du read. Il vise à supprimer les reads dont la taille est trop faible, ce qui rend difficile les analyses effectuées par la suite, comme le mapping. Un read de faible taille peut être par exemple causé par les différentes suppressions décrites ci-après. Enfin, il est possible de filtrer les reads de faible complexité, il s'agit par exemple de reads où toutes les nucléotides sont identiques.

Suppression des adaptateurs

Le processus de suppression des séquences d'adaptateurs, appelé «read trimming» ou «clipping», est une étape initiale cruciale dans l'analyse des données de séquençage NGSs. Les adaptateurs sont nécessaires lors de la préparation des bibliothèques de séquençage, mais ils peuvent contaminer les séquences si les fragments d'ADN sont plus courts que le nombre de cycles de séquençage, le séquençage se poursuivant sur l'adapteur. Cette pollution des reads n'est pas idéale pour un alignement, c'est pour cela que des outils existent afin de supprimer ces adaptateurs des séquences.

Suppression des bases de basse qualité

La qualité des reads n'est en général pas homogène. Par exemple, sur la plateforme Illumina la qualité de lecture des bases à tendance à décroître à la fin du read. Celà peut être dû à une erreur de *phasing*, où un fragment d'ADN manque un cycle de séquençage, causant son marqueur à être en décalage avec les autres fragments du cluster. Cette erreur de *phasing* vient alors affaiblir le consensus dans le signal du cluster et fait baisser le score de qualité. Ce phénomène est bien entendu plus probable vers la fin de la lecture des reads [157].

Il est possible de couper les reads au niveau de ces parties de faible qualité afin de conserver la partie du read la plus fiable. Pour cela, il est possible de calculer le score moyen à l'aide d'une fenetre glissante sur le read, et supprimer l'extrémité de faible qualité.

Suppression polyG

On appelle polyG une succession de guanine située en fin de reads. Cela est expliqué par le fait que certaines plateforme de séquençage utilisent l'absence de signal lumineux pour représenter la guanine. L'intensité du signal lumineux diminuant en fin de read, les nucléotides deviennent reconnus incorrectement comme des guanines. Il est par conséquent souhaitable de supprimer ces fins de séquence [127].

2.7.4 Alignement

Pour étudier des séquences d'ADN, il est bien souvent nécessaire d'effectuer la comparaison entre différents fragments. Même pour la simple tache de recherche d'une séquence dans un génome, on veut bien souvent effectuer cette recherche de manière approximative à cause des nombreuses variations structurales que peut comporter l'ADN, qui rendrait une recherche exacte inutile.

L'alignement est une manière de représenter deux ou plusieurs séquences d'ADN de manière à mettre en évidence leur points communs et leur différences. On distingue l'alignement global (Listing 2.1), qui couvre les séquences dans leur intégralité, et l'alignement local (Listing 2.2) qui se concentre sur des segments spécifiques où la similarité est la plus élevée, en négligeant les autres parties des séquences. L'alignement global est souvent utilisé lorsque les séquences à comparer sont de longueur similaire et présentent une grande similarité.

Listing 2.1 – Exemple d'alignement global

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
   ||||||||||     |||||||  ||||||||||||| |||||||
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
```

Listing 2.2 – Exemple d'alignement local

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
            ||||  ||||||  ||||||||||||||
       5' TACTCACGGATGAGGTACTTTAGAGGC 3'
```

Quand les deux séquences ne sont pas identiques, on parle de «gap» et de «mismatch». Un gap correspond à un indel (une insertion ou une deletion). Un mismatch correspond à une différence entre les nucléotides d'une séquence par rapport à l'autre. L'algorithme d'alignement va chercher à minimiser les gaps et mismatches.

Listing 2.3 – Exemple de mismatch au niveau du 7ème nucléotide

```
5' ACTACTAGATT 3'
   ||||||  ||||
5' ACTACTTGATT 3'
```

Listing 2.4 – Exemple de gap de taille 2 au niveau du 6ème et 7ème nucléotide

```
5' ACTACTAGATT 3'
   |||||   ||||
5' ACTAC--GATT 3'
```

Le format Compact Idiosyncratic Gapped Alignment Report (CIGAR) permet de représenter un alignement d'une séquence sur une référence en codant une séquence d'événements (par exemple les mismatches, insertions, suppressions). Par exemple pour la séquence Listing 2.5, le CIGAR «2S5M2D2M» correspond à :

- 2S : 2 *soft clipping* (peut être des mismatches, ou une séquence plus longue que la séquence correspondante). On parle de *clipping* quand les extremités du read ne sont pas alignées.
- 5M : 5 matches ou mismatches
- 2D : 2 délétions
- 2M : 2 matches ou mismatches

Listing 2.5 – Exemple d'alignement correspondant à la chaîne CIGAR 2S5M2D2M

```
5' GTCGTAGAATA 3'
     |||||   ||
5' CACGTAG--TA 3'
```

Alignement pairwise

Les méthodes d'alignement de séquences *pairwise* sont utilisées pour trouver les alignements locaux ou globaux les mieux adaptés de deux séquences interrogées. Les alignements *pairwise* ne peuvent être utilisés qu'entre deux séquences à la fois, mais ils sont efficaces à calculer et sont souvent utilisés pour des méthodes qui ne nécessitent pas une extrême

précision (comme la recherche dans une base de données de séquences ayant une grande similitude avec une requête).

Par exemple, l'outil BLAST (basic local alignment search tool), permet de rechercher dans une base de données de séquences d'ADN les séquences similaires à une séquence «requête», et identifier les séquences de la base de données qui ressemblent à la séquence «requête» au-dessus d'un certain seuil.

Alignement de multiples séquences

L'alignement de multiples séquences consiste à aligner 3 séquences ou plus.

Les alignement de multiples séquences nécessitent des méthodologies plus sophistiquées que les alignements par paires, car ils sont plus complexes sur le plan computationnel. La plupart des programmes d'alignement de multiples séquences utilisent des méthodes heuristiques plutôt que la recherche d'un optimum global, car identifier l'alignement optimal de plus de quelques séquences de longueurs modérées est excessivement coûteux en termes de calcul.

2.7.5 Mapping

FIGURE 2.27 – Visualisation d'un mapping sous le logiciel Tablet [184]

Il est difficile d'obtenir une vue compréhensible du génome uniquement à partir de reads des séquenceurs, qui sont généralement courts dans le cadre du séquençage Illumina. Le caractère clonal du MTBC, ainsi que la structure relativement commune et bien conservée des différents génomes du MTBC, rendent particulièrement adaptée la méthode du «mapping» [180]. Cette méthode se base sur la comparaison avec la séquence d'une souche de référence, contrairement aux méthodes «*de novo*» qui consistent à reconstruire la séquence du génome sans référence.

Pour procéder à un mapping il est nécessaire d'obtenir une séquence de référence. Il s'agit d'une séquence complète d'un génome très proche de celui séquencé. Pour le MTBC, on utilise couramment $H_{37}Rv$. Le programme de mapping va aligner chaque read du séquenceur à sa position correspondante sur le génome de référence. Il ne s'agit pas d'un placement exact car le génome peut contenir des répétitions, donc la séquence aura plusieurs endroits possible où être placée. Le génome séquencé peut aussi contenir des séquences non présentes dans la référence, et ces reads ne pourront alors pas être placés.

Un programme de mapping couramment utilisé est Burrows-Wheeler Alignment (BWA), il permet d'aligner un très grand nombre de séquences sur une séquence de référence. Ce type de programme a comme sortie un fichier sequence alignment map (SAM) qui est un format texte permettant de stocker des séquences d'ADN alignées sur une séquence de référence. Le format binary alignment map (BAM) est la représentation binaire compressée sans perte du format SAM. Le format Compressed Reference-oriented Alignment Map (CRAM) va plus loin dans la compression mais nécessite un accès à la séquence de référence lors de la lecture. Une visualisation d'un mapping est donné figure 2.27.

Dans les cas de données de séquençage paired-end, la qualité du mapping se trouve améliorée. La distance entre la paire de séquence étant connue, cette information peut ainsi être utilisée pour un mapping plus précis au niveau des régions répétitives. Cela permet un alignement de bien meilleure qualité car si un read peut être mappé à plusieurs endroits d'une référence, il est beaucoup moins probable pour une paire de reads de faire de même (figure 2.28).

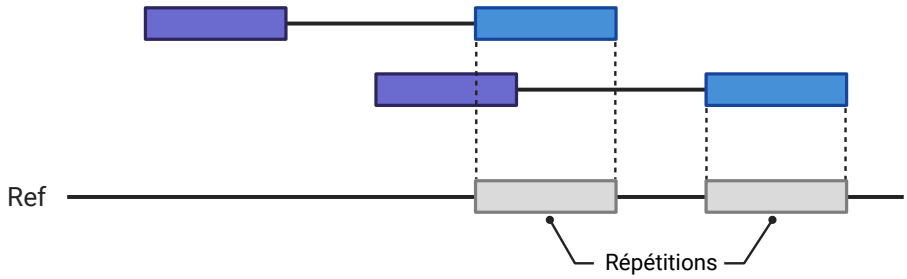


FIGURE 2.28 – Un mapping plus précis de reads (en bleu) de régions répétées a été possible grâce aux seconds reads qui eux se situent dans une région unique du génome.

Une fois le mapping effectué, il est possible de déterminer plusieurs statistiques afin d'évaluer la qualité du mapping. La mesure principale est la couverture («coverage»), qui correspond au pourcentage de bases du génome de référence couvertes par des reads. Une couverture très faible pourrait par exemple indiquer que les reads mappés sont de très mauvaise qualité ou s'ils correspondent à une espèce différente de la référence. Une autre mesure importante est la profondeur moyenne («mean depth»), cela correspond à la moyenne du nombre de reads mappés sur chaque base du génome de référence. Plus la valeur est élevée, plus il sera possible de déterminer avec confiance les différentes caractéristiques du génome, mais les calculs seront aussi plus coûteux en temps. Il existe d'autres mesures, dont le calcul est propre au logiciel de mapping utilisé, comme par exemple la qualité du mapping. Il est par exemple possible d'utiliser l'outil SAMtools pour effectuer ces calculs sur les fichiers SAM [67].

2.7.6 Souches de référence du MTBC

La séquence de référence la plus courrament utilisée pour l'analyse des génomes du MTBC est une séquence de la souche $H_{37}Rv$. H_{37} a été isolée pour la première fois par

Edward R. Baldwin en 1905, sur un homme de 19 ans atteint de tuberculose pulmonaire chronique. C'est ensuite que le variant «Rv» fut obtenu, «R» pour «rough colony structure» et «v» pour «virulent», par culture par Steenken [261] en 1935. La séquence complète de son génome a été déterminée par Cole et al. [41] en 1998. Elle est actuellement identifiée au NCBI sous le numéro `NC_000962.3`.

Une autre souche de référence est *Mycobacterium tuberculosis 18b* [29], qui est basée sur une souche non virulente, ainsi que AF2122/97 pour le variant *bovis* [97].

2.7.7 Assemblage *de novo*

L'assemblage *de novo* consiste à reconstruire la séquence d'un génome à partir de reads mais sans séquence de référence. Les programmes d'assemblage fonctionnent en rassemblant les reads se chevauchant pour construire des séquences génomiques continues, appelées «contigs», dont l'ordre des bases est connu avec un haut niveau de confiance. Les contigs sont ensuite assemblés en «scaffolds» afin de tenter de reconstituer la séquence complète du génome. Les reads longs sont à privilégier pour les approches d'assemblage *de novo*, en particulier lorsque l'on vise à obtenir des génomes complètement circularisés et à générer de nouvelles références génomiques [180].

SPAdes (St. Petersburg genome assembler) [13, 214] est un algorithme d'assemblage de génomes conçu pour les ensembles de données bactériennes monocellulaires et multicellulaires. SPAdes fonctionne avec des séquences issues de PacBio, Oxford Nanopore, single-end et paired-end d'Illumina. SPAdes est intégré dans le pipeline Shovill qui modifie les étapes avant et après l'étape d'assemblage principale pour obtenir des résultats similaires en moins de temps, l'assemblage étant bien souvent très long. Le pipeline Shovill ajoute par exemple une phase de sampling aléatoire au début du pipeline afin de réduire la quantité de reads à assembler dans le cas où le nombre total de reads est disproportionné par rapport à la taille du génome.

2.7.8 Variant calling

Le variant calling est le processus par lequel on identifie des SNPs et des indels à partir de données de séquençage. Une des principales difficultés est le fait que les données de séquençage peuvent contenir du bruit, il est ainsi parfois difficile de distinguer les vraies mutations des faux positifs.

Variant calling à partir d'un mapping

Lorsqu'un génome de référence est à disposition, une des principales méthodes est de commencer par un mapping des reads sur la référence. La qualité des résultats sera par conséquent liée à la qualité de ce mapping. Il existe des méthodes pour améliorer le mapping afin de mieux l'adapter au calling, comme par exemple *samclip* [236] qui supprime les reads fortement clippés.

Une première méthode de variant calling à partir d'un mapping pourrait être simplement de comparer une à une les nucléotides de la séquence de référence avec les nucléotides des reads mappés. Cette méthode peut aboutir à des résultats paradoxaux, par exemple dans le Listing 2.6, deux mutations seraient détectées sur la séquence de référence alors qu'aucun des reads appuie ce résultat car aucun read n'a deux mutations.

Listing 2.6 – Exemple de deux reads mappés sur une même référence

```
R1.   5' ACTACTAGATT 3
R2.   5' ACTACTTGAAT 3'
         |||||| || |
Ref.  5' ACTACTTGATT 3'
```

Une méthode alternative, qui est maintenant communément utilisée, est la méthode par haplotype. Avec cette méthode on détermine les variations en évaluant des groupes de mutations, et non des positions individuelles. Autrement dit, ce n'est plus l'alignement qui est directement utilisé mais les reads eux-mêmes afin de trouver un consensus. Les méthodes par haplotype commencent par utiliser le mapping pour trouver des régions contenant de nombreuses variations (régions dites polymorphiques). Une fois les régions trouvées, ce sont les reads qui sont utilisés pour déterminer l'haplotype le plus probable. Il existe différentes méthodes statistiques pour le déterminer dont, par exemple, *freebayes* [98].

Variant calling ciblé

Il existe aussi des méthodes sans mapping sur une séquence de référence. Le mapping est intensif en calcul, générant de grands fichiers de données intermédiaires. Il n'est parfois nécessaire de seulement déterminer la présente ou non d'un ensemble de variations à partir de reads, un mapping complet n'est alors pas nécessaire.

KvarQ [262] est un exemple d'outil qui scanne directement les fichiers fastq des séquences de génomes bactériens pour des variants connus, tels que les SNPs, contournant ainsi le besoin de mapper tous les reads sur un génome de référence. Dans une première étape, KvarQ extrait les flancs (les séquences de chaque côté d'une région) entourant la position sur laquelle il est souhaité de déterminer la présence ou non d'un SNP. Ces séquences flancs sont ensuite recherchées directement dans les séquences des fichiers fastq. Les séquences de faible qualité ou les correspondances avec un nombre de reads trop faibles sont ignorées.

Cette méthode possède la limitation importante qu'elle n'est pas adaptée pour déterminer tous les variants de données de séquençage par rapport à un génome de référence. En effet, il est d'abord nécessaire de déterminer une liste de variants spécifiques à rechercher. Cependant, dans le cadre de recherche de mutations liés à la résistance aux antibiotiques qui sont déjà connus, cette méthode permet d'éviter des calculs complexes via un mapping.

Le chapitre d'introduction se conclut, ayant couvert l'ensemble des connaissances et des outils pertinents pour la compréhension des travaux qui vont être maintenant présentés. Le prochain chapitre abordera les lacunes identifiées dans les méthodes d'étude de *M. tuberculosis* et justifiera l'introduction d'une nouvelle plateforme d'analyse, TB-annotator, qui constitue une première contribution à ce domaine de recherche.

Chapitre 3

TB-annotator : une plateforme pour l'analyse à large échelle des génomes du MTBC

Le présent chapitre constitue une version française détaillée de l'article introduisant la plateforme TB-annotator. La version condensée anglaise est incluse en annexe A.

3.1 Introduction

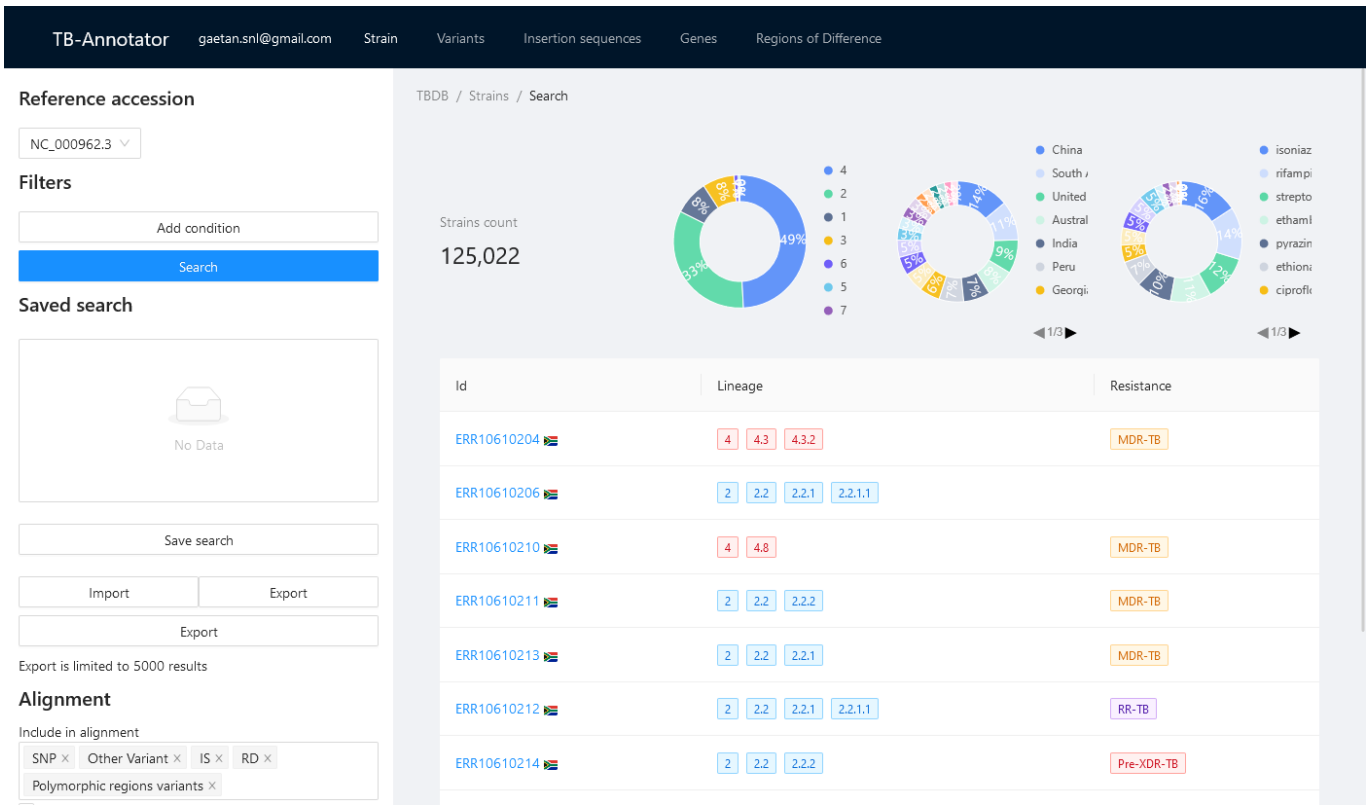


FIGURE 3.1 – Page d'accueil de TB-Annotator. Sur la gauche le panneau de filtrage et sur la droite la liste des souches correspondantes. Les statistiques sont calculées dynamiquement en fonction des filtres actifs.

Le coût du séquençage a chuté de manière spectaculaire au cours des 25 dernières années, rendant l'acquisition de nouvelles séquences génomiques courante et abordable. Les technologies NGS ont permis de collecter plus de 160 000 séquences de génomes MTBC

disponibles publiquement sur la base de données NCBI SRA [2], le référentiel bien connu pour les données de séquençage à haut débit. Une telle base de données constitue une source d'information essentielle pour l'étude de l'évolution de la tuberculose, mais les outils manquent évidemment pour l'exploiter pleinement. Par exemple, on peut constater le manque de standardisation des pipelines d'analyse WGS et le besoin de bases de données pour partager les données WGS à un niveau mondial [180].

L'objectif de cette nouvelle plateforme n'est pas d'offrir un énième pipeline reproduisant ce qui est réalisé par d'autres pipelines existants, mais de proposer un nouveau paradigme dans les systèmes d'analyse de la tuberculose. Le besoin d'un tel outil est déterminé par les observations suivantes.

Tout d'abord, les outils existants se concentrent principalement sur des sujets spécifiques comme la prédiction de la résistance ou le génotypage des isolats bactériens. Leur objectif n'est pas de fournir une approche globale et complète basée sur l'intégralité de l'information génomique disponible, mais quelque chose de spécifique aux données fournies par l'utilisateur. Par exemple, certains pipelines ont été développés pour prédire la résistance potentielle aux médicaments, comme les reconnus et largement utilisés TB-profiler et/ou Phyresse [213, 85]. Ces pipelines n'ont pas d'outils pour explorer et analyser les caractéristiques génomiques à grande échelle. Ils se concentrent généralement sur un type unique de marqueurs, et ils peuvent être dépourvus d'outils d'analyse globale. Certains pipelines utilisent une multitude d'outils, difficiles à maintenir, difficiles à exécuter, et ne permettent pas de passer à l'échelle des génomes complets de MTBC.

Aussi, un certain nombre d'outils ne sont pas, par construction et raison d'être, spécifiques à *M. tuberculosis*. Ils gagnent donc en généralité ce qu'ils perdent en spécificité. S'ils peuvent être appliqués de manière non spécifique à une variété de bactéries, ils ne peuvent en revanche pas tirer parti de décennies de connaissances spécifiques accumulées sur les caractéristiques génomiques du MTBC. Chaque pipeline produit des résultats différents dans un format incompatible, ce qui rend la construction d'un pipeline d'analyse global, intégrant plusieurs outils complémentaires, plutôt complexe à réaliser. Aussi, il est souvent difficile de comparer des souches privées avec des génomes déjà disponibles publiquement. Enfin, mettre en place des workflows entièrement intégrés à partir de zéro, permettant une analyse précise et significative des données NGS provenant de souches cliniques du MTBC, nécessite encore une expertise en programmation et un personnel bioinformatique formé. Cela contraint l'application de l'analyse NGS de MTBC à des laboratoires spécialisés, conduit à une grande diversité de pipelines d'analyse avec des solutions spécifiques à chaque groupe, ce qui complique sérieusement la comparaison standardisée des résultats.

C'est pour ces raisons que nous avons développé *TB-Annotator*. Cette base de données est à la même échelle massive que celle des SRA, mais elle intègre en plus la plupart des caractéristiques génomiques connues des génomes MTBC (définitions des marqueurs de lignée, régions de différence, séquences d'insertion...). En tant que telle, elle est la plus grande base de données préanalysée spécifique au complexe *Mycobacterium tuberculosis*. De plus, cette base de données est dotée d'une interface ergonomique et originale, permettant de réaliser des requêtes avancées ou de créer des filtres complexes, et d'explorer avec une profondeur encore inégalée, tout arbre phylogénétique spécifique du choix de l'utilisateur. Une telle masse de données exploitée par ces capacités d'analyse avancées permet de s'affranchir d'une partie des biais de sélection. Chaque étude s'effectue sur l'intégralité des données génomiques dont nous disposition, et non plus sur des *BioProject* précis. Comme il sera montré dans les sections suivantes, *TB-Annotator* a déjà permis de revisiter un certain nombre de connaissances sur les lignées de la tuberculose [105, 238].

À notre connaissance, une telle plateforme capable de traiter une quantité massive de données n'est pas encore disponible pour les génomes du MTBC. Concrètement, cette grande quantité de données de diversité génomique et spatio-temporelle encore inexploitées pourrait constituer dans un avenir proche une mine d'or de connaissances encore cachées, permettant de réaliser une plongée archéologique dans le passé de l'histoire des maladies

infectieuses. Avec plus de 160 000 SRA disponibles, et en constante augmentation, le domaine est assez mature, et une partie du travail consistera à être capable de discriminer quelles sont les informations qui amélioreront réellement notre compréhension globale de la pandémie de tuberculose.

Ce faisant, l'un des objectifs finaux de ce projet évolutif pourrait être de concevoir pour la tuberculose un projet similaire à celui qui a été lancé juste avant la crise sanitaire du Covid-19, dont l'ambition était de réaliser un suivi en temps réel des agents pathogènes [112].

On peut distinguer deux composants principaux de *TB-Annotator*. Premièrement, le pipeline qui permet une analyse massive et uniforme de tous les génomes publics du MTBC, ainsi que de génomes privés de manière optionnelle. Il s'agit d'extraire les différents marqueurs génomiques bruts des séquençages publics. Le scond composant, la plateforme d'analyse, sert à exploiter ces données brutes en les enrichissant puis en permettant d'effectuer des recherches et calculs en temps réel sur l'ensemble des données.

3.2 Pipeline

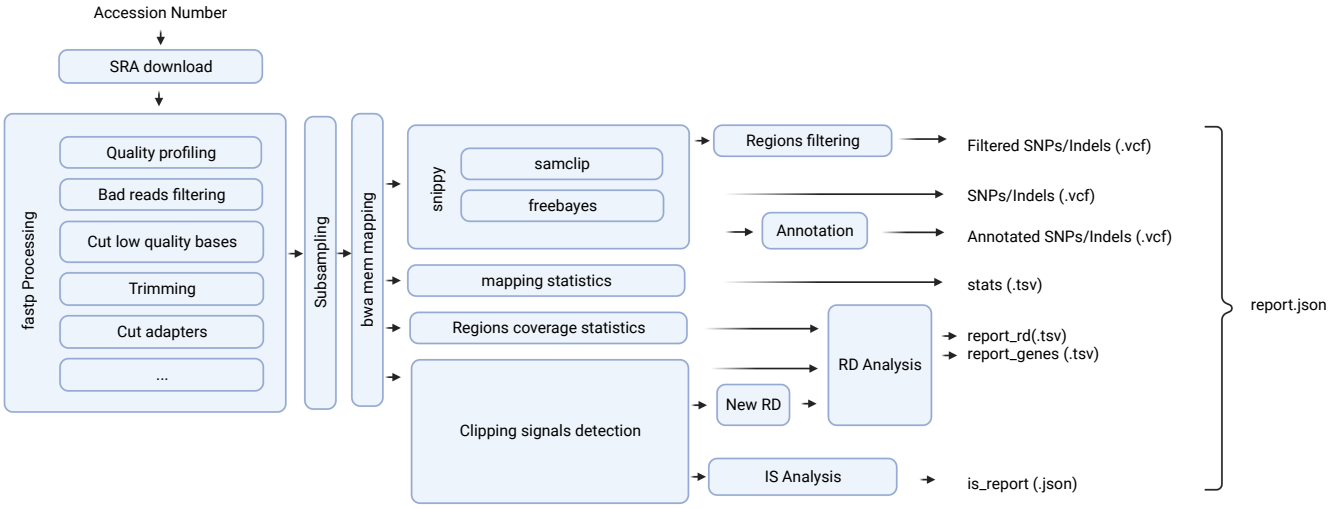


FIGURE 3.2 – Architecture générale du pipeline

3.2.1 Architecture générale

Tout d'abord, le pipeline utilise Snakemake [190] comme moteur d'exécution, qui est lui même basé sur l'écosystème Python. Snakemake est conçu pour construire des pipelines reproductibles et évolutifs en tirant parti d'une définition déclarative des workflows et de plateformes d'exécution modernes telles que le cloud (par exemple, Kubernetes [153]). La reproductibilité est l'un des défauts majeurs de beaucoup d'outils bioinformatiques, où il est difficile de réexécuter l'outil dans un environnement indépendant. Le fait d'utiliser un moteur d'exécution existant, associé à une gestion des dépendances par l'outil Conda, permet de remédier à ce problème. Les étapes d'exécution sont organisées sous forme de graphe puis exécutées en fonction des dépendances de données entre chaque étape. Ce moteur de workflow a été choisi plutôt que des moteurs de workflow cloud-native tels que Argo Workflows [10] pour sa facilité d'utilisation, sa facilité de déploiement sur des ordinateurs personnels et l'intégration facilitée avec les outils bioinformatiques courants. Il facilite aussi la sauvegarde des états intermédiaires du pipeline, permettant ainsi d'interrompre et de reprendre l'exécution.

Le pipeline commence par une liste de numéros d'accession fournie par l'utilisateur au format TSV. Les séquences sont ensuite téléchargées par le pipeline depuis la base de données SRA [159], avec fastq-dump, sous forme de fichiers FASTQ en single-end ou paired-end. Cette première étape du pipeline est régulée pour réduire le nombre de connexions simultanées au NCBI. Les scores des fichiers FASTQ sont utilisés dans les étapes de prétraitement pour évaluer la qualité des reads, par conséquent, le format FASTQ est le seul format d'entrée pris en charge.

Pour les séquences non disponibles publiquement, le pipeline accepte les fichiers FASTQ compressés. Par convention, les séquences provenant de sources privées reçoivent un numéro d'identification personnalisé commençant par le préfixe *CUS*, pour «custom» (par exemple CUS000001). Le pipeline est préconfiguré pour détecter ce schéma de dénomination et pour analyser les séquences *CUS* en même temps que les séquences SRA. Ce système de numérotation permet d'anonymiser par défaut chaque souche soumise au pipeline. En effet, seul le laboratoire ayant fourni les souches associe ce numéro à des métadonnées, le partage des métadonnées avec TB-Annotator est ainsi optionnel.

Le DAG (directed acyclic graph, i.e. graphe sans cycle orienté) du pipeline et les calculs sont organisés pour maximiser la réutilisation des données et les améliorations incrémentales. Cela permet d'effectuer de nouvelles analyses et calculs sans réexécuter des étapes de calcul coûteuses. Cette organisation des étapes optimisant le temps de calcul est une caractéristique unique du pipeline *TB-Annotator*, qui explique le faible nombre d'étapes et d'outils par rapport aux pipelines existants.

Préparation des reads

La couverture du génome est estimée en fonction du nombre de reads dans le fichier FASTQ. Pour garantir que la vitesse du pipeline reste constante, le fichier FASTQ est échantillonné aléatoirement à une couverture génomique estimée configurable en utilisant SeqKit [243].

Fastp [33], un outil tout-en-un et haute performance, est utilisé pour le contrôle de qualité des reads et leur prétraitement. Dans la partie prétraitement, les adaptateurs utilisés lors de la préparation de la bibliothèque pendant le séquençage sont retirés. Les bases de faible qualité aux extrémités des reads sont coupées et les reads avec trop de bases de faible qualité sont supprimés. Pour les reads paired-end qui se chevauchent, ils sont corrigés en fonction du score de qualité s'ils ne sont pas un parfait reverse-complement. Certaines corrections avancées sont également effectuées, comme le *clipping* des polyG et le prétraitement des unique molecular identifiers (UMI) [148].

Après le contrôle de qualité et le prétraitement des reads, un rapport est généré aux formats html et json. Le rapport contient des informations et des statistiques sur les différentes parties des reads filtrés, le contenu en bases et les taux de duplication. Ce rapport peut être utilisé pour évaluer le processus de séquençage et détecter des problèmes tels que des reads contaminés, des biais dans le taux de bases et des séquences surreprésentées. Les fichiers FASTQ originaux sont supprimés après ce processus pour réduire l'utilisation de l'espace disque.

Cette phase est séparée en deux chemins dans le graphe, selon que les reads sources sont de type paired-end ou single-end. Les étapes réalisées sont identiques, mais les fichiers utilisés sont différents. Dans le cas du single-end un seul fichier est utilisé, et dans le cas du paired-end deux fichiers sont utilisés. Après l'étape suivante, il n'y aura plus de différence selon le type de read.

Mapping des reads

Les reads traités sont mappés sur un génome de référence (`NC_000962.3` par défaut pour MTBC) en utilisant BWA-MEM [162]. BWA-MEM est l'un des mappers les plus rapides [167], tout en étant proche de Novoalign en termes de précision [118]. Les reads

dupliqués sont marqués à l'aide de SAMTOOLS [161] pour les outils suivants dans le pipeline, y compris freebayes, afin de prévenir les biais dans le variant calling dus aux duplications PCR qui pourraient amener l'algorithme à identifier une erreur lors de l'amplification comme un véritable variant [78].

L'alignement des séquences est sauvegardé au format CRAM, un format compressé basé sur la séquence de référence [124]. Le fichier résultant est suffisamment petit pour être conservé sur un disque dur à faible coût, et permet d'effectuer des analyses ultérieures ou d'exécuter un pipeline mis à jour à l'avenir sans avoir à exécuter une seconde fois l'étape de téléchargement et de prétraitement. Cette optimisation est rendue possible grâce à la capacité de Snakemake à continuer des workflows partiellement exécutés. Le temps de calcul est réduit en utilisant l'alignement des séquences comme base pour toutes les étapes suivantes du pipeline, au lieu d'exécuter des analyses sur les reads bruts. Un index du fichier CRAM est également généré et sera conservé avec le fichier CRAM.

Enfin, des statistiques sont calculées avec SAMTOOLS [161] en utilisant le fichier d'alignement des séquences, comme le nombre de reads alignés, le nombre de bases couvertes avec une profondeur ≥ 1, le pourcentage de bases couvertes, la profondeur moyenne de couverture et la qualité moyenne des bases. Ces informations peuvent être utilisées pour évaluer la qualité de l'échantillon et celle du séquençage. Ces informations seront également utilisées dans le reste du pipeline pour des analyses statistiques.

3.2.2 Détection des mutations

Variant calling

Snippy [235] est utilisé pour le variant calling, qui est l'un des pipelines les plus performants pour les génomes bactériens [28]. Snippy prend en entrée les reads déjà alignés au format CRAM compressé au lieu de les aligner (par défaut avec bwa-mem aussi) au format BAM. Un réalignement local n'a pas été effectué avant le variant calling car il avait un faible impact avec les outils utilisés [28].

Snippy commence par l'utilisation de samclip [236] sur les reads alignés pour éviter les faux SNP près des variations structurales. Il fonctionne en supprimant les reads clipped à l'intérieur des contigs. L'alignement résultant contient principalement des reads longs alignés, fournissant ainsi une plus grande confiance dans le variant calling. Les variants sont ensuite détectés avec Freebayes [98]. On considère un site quand la profondeur de reads y est au minimum de 10, que la probabilité d'erreur issue de la qualité est inférieure à 5%, que le read est mappé en un endroit unique et que 90% des reads indiquent la mutation. Le résultat final de cette étape est un fichier Variant Call Format (VCF).

La variant calling est un processus imparfait, où un certain nombre de faux positifs et de faux négatifs seront créés. Les variants ne sont donc pas à considérer comme une information binaire. Le fichier VCF contient des informations permettant d'évaluer la qualité de chaque information, notamment la profondeur au niveau du site du variant, le nombres de reads avec la base de référence, le nombre de reads avec la mutation, et leur qualité respective.

Certaines régions des génomes du MTBC sont connues pour être difficiles à mapper, y compris les éléments mobiles, les gènes Proline-Glutamate (PE)/Proline-Proline-Glutamate (PPE) et les régions répétitives. Ces régions contiennent généralement un certain nombre de reads soft-clipped qui sont supprimés par samclip, mais certains d'entre eux sont conservés, par exemple les reads correspondant aux séquences d'insertion mappées aux régions IS de la référence MTBC. Ces régions problématiques sont supprimées dans plusieurs études [175]. Un fichier VCF supplémentaire, filtré avec *bedtools intersect*, est produit dans lequel les variants se trouvant dans ces régions sont supprimés.

Nous identifions les variants en utilisant leur notation SPDI [122], ce qui permet d'avoir un identifiant unique global pour chaque variant. SPDI représente tous les variants comme une séquence de quatre opérations : aller à la position 0 de la séquence S,

avancer de P nucléotides, supprimer D nucléotides et insérer les nucléotides I. Par exemple, `NC_000962.3:10:1:G` correspond à un SNP où le 11ème nucléotide (10ème en base 0) de `NC_000962.3` devient G. La représentation alternative dans laquelle le champ Deletion est une chaîne contenant la séquence littérale à supprimer [122] est utilisée pour éviter le besoin de recourir à la séquence de référence lors de la lecture du SPDI. On utilisera donc par exemple `NC_000962.3:10:A:G`.

Annotation

La liste des variants obtenue fournit uniquement une information génomique, mais cette information peut être utilisée pour déduire les effets qui en résultent. Par exemple, les changements d'acides aminés peuvent être déduits grâce aux variants ainsi qu'un génome de référence annoté.

Pour cette étape on utilise SnpEff [37], un outil d'annotation de variants et de prédiction d'effets. Une base de données de référence est créée en utilisant la séquence de référence `NC_000962.3` de $H_{37}Rv$ ainsi que les annotations au format GenBank [160]. Pour chaque souche, chaque variant est ensuite annoté avec SnpEff. Toutes les annotations sont supportées et sont extraites par le pipeline sous le standard «The Sequence Ontology», qui est un vocabulaire contrôlé et structuré pour les différentes annotations génomiques [79]. On retrouve des annotations avancées comme la détection d'un variant changeant le cadre de lecture, changeant un codon d'initiation ou de terminaison. Mais on retrouve aussi des annotations plus globales, comme les variants synonymes et non synonymes, et ce sont ces annotations issues de SnpEff qui seront utilisées directement dans la plateforme.

SnpEff est aussi utilisé pour obtenir les notations HGVS des variants. HGVS étant un standard pour décrire les variants des séquences d'ADN mais aussi des variants dans les protéines [73]. Cette notation permet donc, par exemple, de décrire un variant sous forme de changement d'acide aminé au lieu d'un changement de nucléotide dans la séquence d'ADN. C'est pour cela que ces notations sont particulièrement utilisées dans les bases de données décrivant les variants marqueurs de résistance, où la résistance à un médicament est issue d'un changement d'acide aminé.

3.2.3 Méthodes pour la détection des variations structurales

La présence de variations structurales dans le génome analysé entraîne des comportements caractéristiques lors du mapping des reads. Nous développons des techniques basées sur deux de ces comportements : les variations de la profondeur de lecture et l'apparition de signaux de *clipping*.

Pour la variation de la profondeur, dans le cas d'une diminution, elle signifie que peu de reads ont pu s'aligner dans une région. Une région supprimée, dans le cas où elle est unique sur le génome, va entraîner une diminution à zéro de la profondeur. Dans le cas où la région n'est pas unique, il n'y aura pas de diminution de la profondeur, les reads des régions dupliquées venant combler la région supprimée lors du mapping.

Lorsqu'un read correspond partiellement à la séquence de référence, le mapper supprime généralement la partie non alignée, ce processus est appelé *clipping*. Il existe deux types de *clipping* : (1) le *hard clipping* où la partie supprimée est effectivement retirée du fichier de mapping résultant, et (2) le *soft clipping* où la séquence supprimée est conservée et sa position est signalée dans la chaîne CIGAR. Par défaut, BWA-mem utilise le *soft clipping* pour l'alignement principal. Un grand nombre de reads clippés aux mêmes positions est généralement un signe d'une variation structurale [272].

Nous appelons *signal de clipping* la position où une quantité configurable de reads est clippée. Un signal de *clipping* peut être du côté gauche ou du côté droit des reads (figure 3.3). Lors de l'étape de détection du signal de *clipping*, un script Python crée la liste des reads clippés, la position du *clipping* gauche et du *clipping* droit, ainsi que la séquence

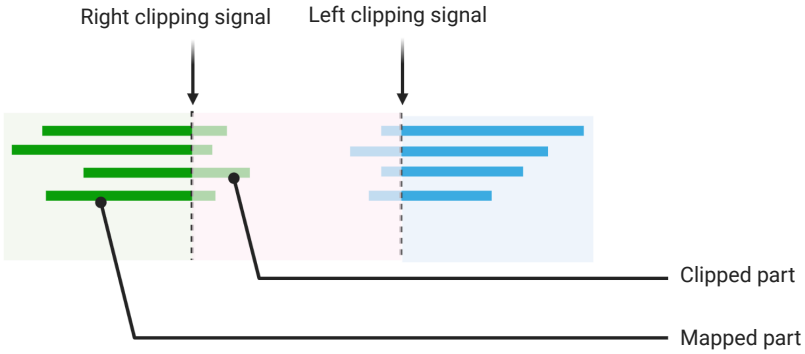


FIGURE 3.3 – Clipping signals

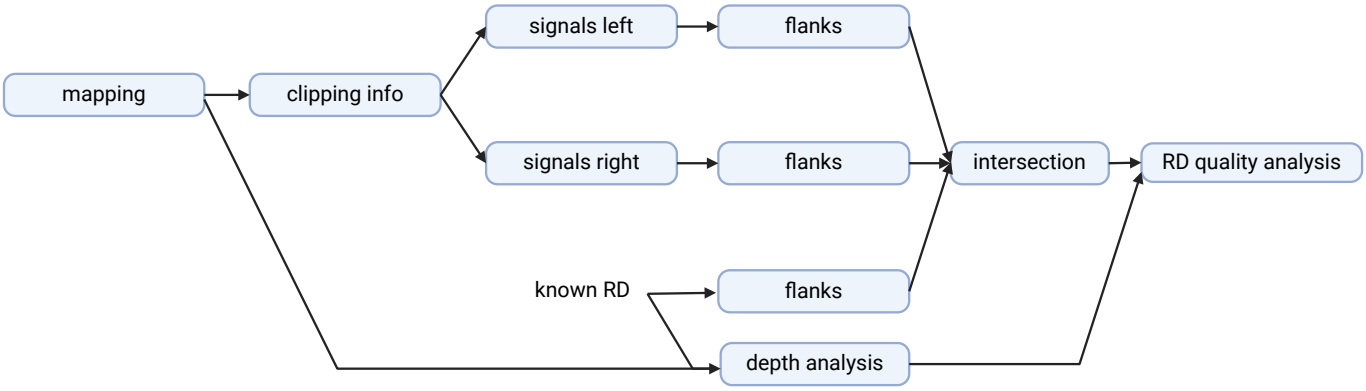


FIGURE 3.4 – Pipeline de données pour la détection des RD. Cette partie du pipeline commence par l'alignement CRAM issu des étapes précédentes. «clipping info» crée la liste des reads clippés, la position du *clipping* gauche et du *clipping* droit, ainsi que la séquence clippée. On considère ensuite comme signal la position où au moins 10 reads clippés s'alignent. On extrait les positions des régions proches des signaux («flanks»), de taille 20 par défaut, et on réalise l'intersection entre ces régions et les flanks internes des régions des RDs. L'intersection de ces régions permet de réaliser le lien entre les signaux et les listes de RDs. Par défaut, une superposition de 50% entre le flank internes d'un RD et du flank du *clipping signal* est nécessaire pour considérer les deux comme liés. C'est lors de la dernière étape que la partie coupée des signaux est alignée sur le côté opposé pour permettre de calculer le score de qualité de la RD.

clippée. La liste des signaux de *clipping* est créée en utilisant *bedtools groupby* [216] sur la sortie du script Python.

Cette étape n'est qu'un prétraitement pour les étapes suivantes. Les informations de *clipping* peuvent être utilisées pour détecter les variations structurales telles que des délétions et des insertions. Et contrairement au cas général de détection des variations structurales, nous pouvons utiliser ces informations pour détecter plus facilement et plus précisément les caractéristiques connues du génome MTBC.

Détection de la présence de RD et gènes connus

Les RD connues ont été collectées manuellement à partir de plus de 20 études, puis organisées et vérifiées. Les positions ont été converties en positions sur le génome de référence $H_{37}Rv$ `NC_000962.3`. Lorsque cette position n'était pas disponible dans l'article original, elle a été déduite en utilisant diverses méthodes comme le mappage de la jonction PCR, le mappage de souches exemples contenant cette délétion, ou l'extraction de la position à partir d'autres références comme *M.bovis*.

Méthode par profondeur Cette étape se concentre sur l'évaluation de la présence/absence de gènes et de grandes RD (les RD beaucoup plus petites que la longueur des

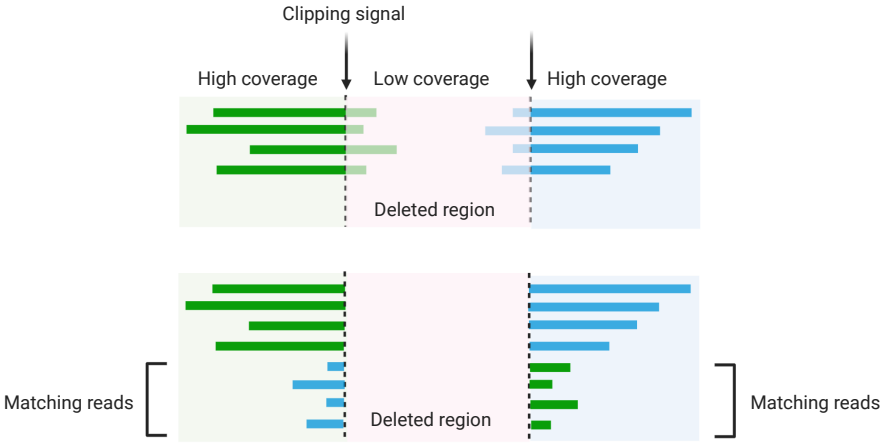


FIGURE 3.5 – Mappage des séquences coupées

reads, comme l'insertion de 7bp dans pks 1/15, sont trouvées par l'étape de variant calling [55]). Un histogramme de profondeur de lecture est calculé sur les reads alignés en utilisant BEDTOOLS [216]. Pour chaque région (RD et gènes), plusieurs statistiques sont calculées comme la moyenne, la médiane, la couverture minimale, maximale et le ratio $\frac{\text{couverture moyenne de la région}}{\text{couverture moyenne globale}}$. La plupart des approches existantes pour la détection des RD *in silico* utilisent ces métriques [81, 317, 61]. Avec cette méthode, nous manquons diverses informations, par exemple : si la région est entièrement supprimée ou non, si la région fait partie d'une délétion plus grande ou est précisément supprimée, s'il est probable que la baisse de profondeur de lecture observée est effectivement liée à une délétion. Pour surmonter certains de ces problèmes, nous calculons le pourcentage de très faible profondeur de lecture dans la région donnée en utilisant un seuil configurable. La valeur résultante peut être utilisée comme indication d'un gène partiellement supprimé par exemple.

Méthode par analyse des clipping signals Pour la deuxième méthode, nous avons développé un algorithme basé sur les *clipping signals* extraits de l'étape précédente. Nous utilisons ces signaux pour trouver ce que nous appelons des *délétions de haute qualité*, qui sont des régions pour lesquelles nous sommes très confiants qu'elles sont des délétions précises (et non des faux positifs dus à une profondeur de lecture plus faible dans la région, par exemple). Les *clipping signals* proches des extrémités de la région sont recherchés en utilisant une marge configurable (figure 3.4). On extrait ensuite l'intégralité des reads clippés des deux signaux de *clipping*. La partie coupée est alignée sur le côté opposé de la région comme décrit dans la figure 3.5. Lorsqu'un nombre configurable de séquences s'alignent précisément des deux côtés, nous marquons la région comme *de haute qualité*. Avec cette approche on confirme donc que tous les reads sur la région sont en réalités les reads de jonction, cohérents avec une deletion.

Le score de qualité est un score compris entre 0 et 1, avec 50% dédié à la qualité du coté droit et les autres 50% pour le coté gauche. Pour chaque séquence alignée sur le côté opposé on calcule un score d'alignement, avec +1 pour match, -1 pour mismatch, -1 pour l'ouverture d'un gap et -0.5 pour son extension. On compte la proportion de séquences dont le ratio entre le score d'alignement et la taille de la séquence est > 0.8. Le score de chaque côté représente 50% du score final.

Les marges sont utilisées lors de la recherche de *clipping signals* principalement en raison de la façon dont les reads sont alignés. Le signal de *clipping* n'est pas toujours proche de la fin des délétions. C'est le cas par exemple lorsque la séquence avant le début de la délétion est égale à la séquence à la fin des régions. Ce cas est illustré dans figure 3.6. Parce que nous savons approximativement où se trouvent les deux signaux de *clipping* pour

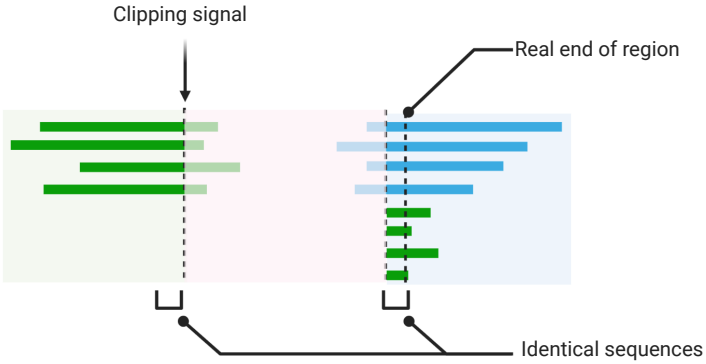


FIGURE 3.6 – Différence entre le signal de clipping et les extrémités des régions supprimées

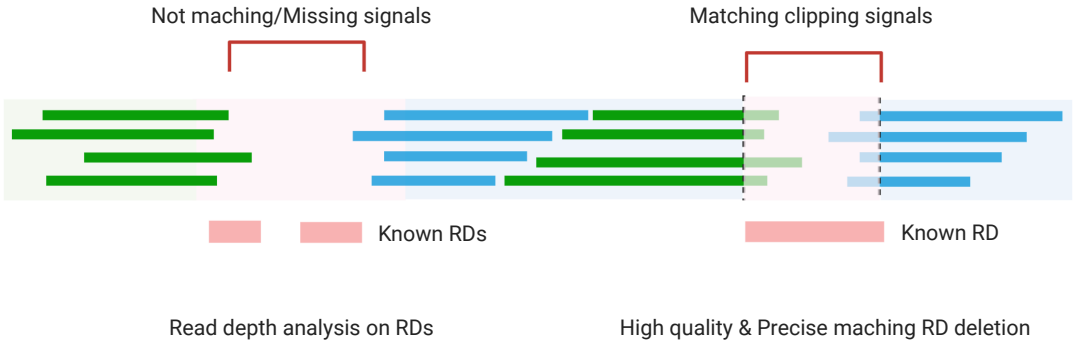


FIGURE 3.7 – Différentes méthodes de détection des délétions de régions

une région donnée, cela diminue le risque de mapping des séquences coupées à une mauvaise position dans le génome. Toutes les régions ne peuvent pas être considérées comment étant de haute qualité, comme les RD tombants dans des délétions plus grandes. Ces régions sont toujours détectées en utilisant la variation de profondeur de lecture comme expliqué précédemment. Ce cas est illustré dans figure 3.7.

Le pipeline est fourni avec deux listes de régions, des gènes et des RD connues, tandis que des régions supplémentaires sont configurables avec des fichiers bed. Pour chaque liste de régions, un rapport TSV est créé, avec toutes les statistiques décrites précédemment et un score indiquant si la région est considérée comme *de haute qualité*. À la fin de cette étape, toutes les régions recherchées ont un score. Le choix final (si une région est supprimée ou non) est laissé à la dernière étape par la création du rapport final. Cela permet de conserver des informations détaillées sur les régions pour une analyse ultérieure.

Détection de nouveaux RDs

Le pipeline permet la détection de RDs qui ne sont pas documentées dans les études existantes. Afin d'assurer la cohérence et la précision des positions détectées dans toutes les souches analysées, cette étape repose également sur les *clipping signals* détectés précédemment.

Les régions de déletion possibles sont recherchées en créant des paires de *clipping signals* consécutifs espacés par un nombre configurable de bases. Chaque paire contient un signal à droite et un signal à gauche. Nous détections donc les patterns qui se produisent en général dans un mapping lors de la délétion d'une région. Un nom est généré pour chacune de ces paires et un fichier similaire à celui des RDs connus est produit. Ces régions sont ensuite analysées en même temps que les régions connues afin d'accroître la rapidité du

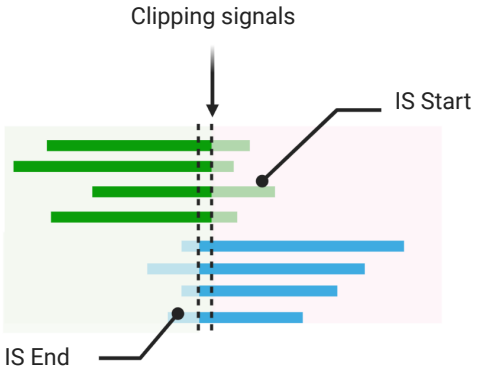


FIGURE 3.8 – Détection des IS avec l'utilisation de clipping signals

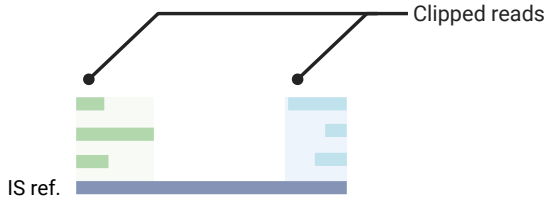


FIGURE 3.9 – Alignement des parties clippées des reads sur la séquence de référence de l'IS

pipeline. Chaque étape n'est donc réalisée qu'une seule fois, seul un script rapide permet de créer les paires potentielles.

Détection des IS

Nous avons collecté des séquences sous forme de fichiers FASTA provenant de 22 séquences d'insertion du MTBC connues (https ://www-is.biotoul.fr/).

Les *clipping signals* détectés lors des étapes précédentes sont utilisés pour détecter les séquences d'insertion. Des paires de *clipping signals* séparés par une petite distance configurable sont recherchées. Pour chaque paire de *clipping signals*, les séquences clippés des reads sont extraites (figure 3.8) et mappées à l'aide de BWA-MEM [162] sur un fichier FASTA contenant toutes les séquences d'insertion connues (figure 3.9).

La séquence d'insertion correspondante est déterminée en analysant le fichier de mapping résultant. Nous trouvons la première séquence où (1) les séquences clippées sont mappées des deux côtés de la séquence d'insertion de référence, (2) les séquences clippées à gauche s'alignent sur la même position, et (3) les séquences clippées à droite s'alignent sur la même position (4) au moins 10 reads sont alignés de chaque cotés. Ce processus garantit que le début et la fin de l'insertion sont cohérents. Cette méthode permet également de détecter l'orientation de l'insertion de l'IS en fonction du côté de la séquence d'insertion de référence où les reads sont mappés. Lorsque la séquence insérée est une sous-séquence d'une IS, nous stockons également le début et la fin de la sous-séquence par rapport à l'IS de référence.

Pour les séquences d'insertion contenues dans la référence, l'algorithme de détection de RD est utilisé pour déterminer si la séquence d'insertion est supprimée. Cette détection est rendue possible grâce à l'algorithme de *clipping* signal. Réaliser cette détection ne serait pas faisable en utilisant des algorithmes de RD reposant sur la profondeur de lecture, car les reads provenant d'autres séquences d'insertion peuvent se mapper sur la région supprimée.

3.2.4 Méthodes pour la reconstruction du CRISPR

On peut distinguer deux résolutions différentes pour la détection du CRISPR. La lecture classique consiste uniquement à déterminer la présence et l'absence des *spacers*. Cependant, cette approche ne permet pas d'explorer de nombreuses autres caractéristiques, par exemple si l'ordre des *spacers* est différent dans une souche ou l'autre. Elle ne permet pas non plus de révéler s'il y eu une duplication d'une partie du locus. Enfin, elle ne fournit aucune information sur la présence d'ISs telles que IS*6110*, ni sur l'existence de SNP dans ses répétitions directes ou dans les espaces. Cela masque des changements potentiellement significatifs sur le plan fonctionnel dans les loci et rend impossible la réalisation d'études évolutives approfondies et plus détaillées à partir du CRISPR [107]. Un second niveau de lecture est la reconstruction complète du CRISPR. Il s'agit ici de reconstruire le locus exact, en prenant en compte aussi l'ordre des espaceurs et les éventuels ISs.

Le premier niveau de lecture est significativement plus rapide à effectuer et est implémenté dans le pipeline afin de fournir un premier niveau d'information. Les reads sont alignés en utilisant BWA-MEM [162] sur les séquences de tous les *spacers* connus. Pour chaque spacer, au lieu de signaler simplement la présence ou l'absence du spacer, nous fournissons des statistiques pour aider à évaluer la probabilité de la présence ou de l'absence du spacer. L'indice de confiance de la détection est basé sur le nombre de reads ayant mappé le spacer, le nombre de bases couvertes et la profondeur moyenne au niveau du spacer. Cette méthode permet aussi de détecter d'éventuels SNPs dans le CRISPR.

Le second niveau de lecture présente de nombreux challenges. La région n'est pas des plus simples à reconstruire, étant parfois par exemple coupée par des insertions d'IS. Dans certains cas où la région est coupée par plusieurs ISs, les méthodes basées sur les short reads ne peuvent pas reconstruire la séquence, car les ISs sont insérés à plusieurs endroits du génome. Il existe la même difficulté sur les régions répétées, la première répétition est reconstruite facilement, mais le nombre exact de répétitions est difficile à déterminer, au mieux une estimation est donnée. Les algorithmes produisent alors, dans le meilleur des cas, des fragments du CRISPR qu'il faut ensuite assembler, mais il existe parfois plusieurs assemblages possibles.

Des outils existent pour la reconstruction du CRISPR en se basant sur des short reads, notamment CRISPRbuilder-TB [107]. Une des techniques permettant à TB-Annotator des analyses à une échelle massive est la réutilisation des résultats des étapes précédentes. Par conséquent, il a été choisi de baser la reconstruction sur une étape de reconstruction réutilisable. Le processus commence par un assemblage classique des reads à l'aide de Shovill [234], lui même basé sur SPAdes [13]. Shovill est un pipeline qui utilise SPAdes comme élément central, mais modifie les étapes avant et après l'étape principale d'assemblage pour obtenir des résultats similaires en moins de temps. Shovill produit un fichier de contigs. On effectue donc bien un assemblage global contrairement à CRISPRbuilder-TB qui construit le graphe de de Bruijn à partir d'un point précis. Le pipeline TB-Annotator effectue un mapping des *spacers*, ISs, DRs et gènes CAS connus sur les contigs. Le mapping est ensuite analysé par un script Python dédié. On annote chaque alignement avec des informations sur le spacer detecté. On reconstruit ensuite un CRISPR en plaçant tout d'abord les *spacers* et DR parfaitement mappés. Ensuite on détecte de chaque côté du contig les éventuels IS et gènes. À ce stade on obtient un fragment de CRISPR avec un haut niveau de confiance sur les éléments placés, car tous les éléments du CRISPR ont été déterminés avec des égalités parfaites par rapport aux séquences connues.La reconstruction du CRISPR contient donc des trous non annotés. En dernière étape, on comble ces espaces avec les séquences partielles les plus proches. Comme dans toutes les autres étapes du pipeline, chaque donnée est annotée avec un indice de confiance. Au final, on obtient des fragments de CRISPR, mais comme précisé précédemment il n'est pas toujours possible d'aller plus loin et d'assembler les fragments de manière déterministe, il s'agit donc de la dernière étape du script dédié au CRISPR. Toutes ces étapes restent coûteuses en temps, et par conséquent sont optionnelles et non exécutées par défaut.

3.2.5 Exécution et résultats du pipeline

Nous exécutons le pipeline avec 128 threads sur un Dell PowerEdge R740 hébergé au Mésocentre de calcul de Franche-Comté. Le système utilise un processeur Intel(R) Xeon(R) Gold 6226R à 16 cœurs fonctionnant à 2,90 GHz avec 128 Go de RAM.

Les SRA du MTBC sont analysés par lots de 15 000, chaque lot prenant environ une semaine pour être analysé. La planification des tâches est gérée par Snakemake en utilisant le planificateur *greedy* pour réduire le temps de calcul du graphe des tâches.

Le résultat de toutes les étapes du pipeline est rassemblé dans un rapport JSON unique contenant :

— la liste des variants trouvés et leurs annotations,
— la liste des gènes manquants,
— la liste des régions de différences manquantes,
— les statistiques de couverture pour chaque gène et RD,
— la liste des séquences d'insertion,
— et le rapport de qualité de Fastp.

Ce rapport contient uniquement les données essentielles à l'analyse. En effet, le volume total de données intermédiaires représente près de 25To, alors que l'ensemble des rapports représente près de 200Go de données.

3.3 Plateforme d'analyse

Le pipeline, une fois exécuté, permet l'obtention de l'ensemble des marqueurs génétiques de l'intégralité des souches du MTBC disponibles publiquement. Ces données sont ainsi standardisées, comparables entre elles et exhaustives de la connaissance génomique actuelle du MTBC. Un tel ensemble de données n'est pas exploitable sans une plateforme d'analyse avancée capable de gérer un tel volume de données. La plateforme d'analyse TB-Annotator est capable de réaliser en temps réel des comparaisons et statistiques sur l'ensemble des caractéristiques génomiques ainsi que des métadonnées disponibles sur l'ensemble des souches disponibles.

On peut distinguer deux composants de la plateforme. D'une part, la base de données qui contient l'ensemble des données analysables, résultat de l'indexation des données en sortie du pipeline. D'autre part, la plateforme de visualisation web qui permet de naviguer, de rechercher et d'exécuter des analyses avancées sur la base de données directement depuis le navigateur de l'utilisateur. L'utilisateur peut effectuer des requêtes rapides sur la plupart des attributs des souches, des variantes, des séquences d'insertion, des gènes et des RD.

3.3.1 Données génomiques brutes et fitrage

Les premiers éléments ajoutés à la base de données sont les données génomiques brutes issues du pipeline. Avec chaque information, différents marqueurs de qualité sont préservés pour être ajoutés eux aussi. Il est nécessaire de trouver un équilibre permettant de préserver le maximum d'information, tout en gardant une taille de base de données acceptable. Pour cela, il n'est indexé que les divers marqueurs et non par exemple les mappings qui apportent peu d'informations quand ils sont identiques à la référence. D'autre part, un filtrage est effectué afin d'exclure certains marqueurs et certaines souches, soit pour des raisons de qualité ou soit pour maintenir la croissance du volume de la base de données constante. Les données les moins pertinentes ne parviennent donc pas à la plateforme d'analyse, mais sont tout de même conservées en sortie du pipeline. Cette séparation permet de changer simplement les critères d'exclusion sans nécessiter une réexecution complète du pipeline.

Les variants sont ajoutés avec les informations de qualité issues du variant calling, c'est-à-dire la profondeur au niveau du site du variant, le nombre de reads avec la base de référence, le nombre de reads avec la mutation, et la qualité de chacun. Les annotations générées par le pipeline à l'aide de SnpEff sont aussi ajoutées dans leur intégralité.

Les RDs sont ajoutés avec les informations issues de la méthode par profondeur (couverture moyenne, couverture médiane, pourcentage de très faible profondeur de lecture et le marqueur de qualité issue de la méthode par analyse des *clipping signals*). Ces RDs sont ensuite filtrés de manière à conserver uniquement ceux ayant un pourcentage de très faible profondeur de lecture supérieur à 90%, un ratio entre la profondeur moyenne de la région et la profondeur moyenne globale inférieur à 10%, et un score de qualité issue du *clipping* d'au moins 80%. Certains RDs qui correspondent à de très courtes délétions sont détectés via le variant calling mais sont ajoutés en tant que RD à la base de données.

Les RDs provenants d'une détection du pipeline et non de listes données utilisent un schéma de nommage *CUS_GS_start_end* avec *start* et *end* le début et la fin de la région. De par la méthode utilisée pour générer la liste de RDs qui est passée à la détection par *clipping*, les RDs CUS contiennent un niveau de bruit important. Ils sont donc filtrés avec le critère le plus exigeant, qui est un score de qualité issue du *clipping* d'au moins 80%.

L'absence des gènes est détectée par la méthode par profondeur car leur délétion est souvent imprécise et parcourt plusieurs gènes consécutifs. Les informations de qualité conservées sont identiques à celles des RDs. Pour le filtrage, on utilise un pourcentage de très faible profondeur de lecture supérieur à 80%, ainsi qu'un ratio entre la profondeur moyenne de la région et la profondeur moyenne globale inférieur à 10%.

Les ISs proviennent de deux méthodes différentes, qui sont les variations structurales pour les ISs de la référence et la détection des ISs pour les nouvelles insertions. On cherche ici la présence des ISs et non leur absence. Pour les ISs de la référence, la méthode par profondeur n'est fiable qu'en cas de très faible profondeur, car des reads provenant d'autres ISs peuvent être mappés dans la région. Les critères sont donc un score de qualité issue du *clipping* d'au plus 80%, et au plus 10% de bases faiblement couvertes ou un ratio entre la profondeur moyenne de la région et la profondeur moyenne globale supérieur à 80%. Pour la détection des nouveaux IS, aucun filtrage supplémentaire n'est nécessaire mais des informations supplémentaires sont ajoutées à la base de données, comme l'orientation et les positions de début et de fin de l'IS inséré. On stocke aussi la méthode de détection de l'IS (via référence ou nouvelle insertion).

Enfin, les souches elles-mêmes sont filtrées lors de l'ajout à la base de données, en se basant sur des critères génomiques. On exclut tout d'abord les souches avec un nombre de variants supérieur à 4500. Il s'agit d'une valeur très élevée qui a pour but d'exclure les souches ayant des problèmes de qualité importants, pouvant faire augmenter considérablement la taille de la base de données. Il arrive parfois que certaines souches contiennent des dizaines de milliers de variants et soient donc très éloignées de la souche de référence utilisée. Les souches avec une profondeur moyenne inférieure à 20 ou avec un pourcentage de bases couvertes inférieur à 80% sont aussi exclues.

3.3.2 Enrichissement des données

Le pipeline produit des données génomiques brutes. Un des objectifs de TB-Annotator est de faire le lien entre l'ensemble des données génomiques disponibles, mais aussi toute la connaissance accumulée sur ces données. Par conséquent, pour chaque caractéristique génomique, nous associons diverses métadonnées collectées au travers de centaines d'articles et de quelques bases de données publiques.

Une base de données de tous les articles associant des variants à des lignées connues a été constituée. Cela permet, entre autres, de catégoriser les variants selon différentes études phylogénétiques. Les variants associés à des résistances aux médicaments (basées sur les données TBDB [223] et WHO [302]) sont aussi ajoutés. Le texte d'introduction de ces études et les informations bibliographiques sont indexées avec les variants ce qui permet aussi d'avoir une visibilité thématique. D'autre part, on exploite la connaissance globale du génome, qui est entièrement annoté, pour associer le variant au gène qui le contient. Enfin, nous faisons le choix de n'exclure aucun variant lors de l'indexation. Comme décrit précédemment, le pipeline produit une sortie additionnelle filtrée pour supprimer les

variants dans les régions difficiles à mapper. Ici nous faisons le choix d'indexer tous les variants, et il sera possible ensuite d'exclure ces régions à l'aide de filtres.

Pour ce qui est des RDs, une opération similaire est effectuée. Une base de données de tous les RDs connus a été constituée (tableau B.1) avec des annotations bibliographiques. Ces annotations sont liées aux RDs lors de l'indexation.

Les séquençages indexés reprennent toutes les informations de leurs marqueurs. Ainsi, un SNP associé à une lignée, permet d'associer aussi une souche à cette lignée. L'annotation des marqueurs permet ainsi qu'annoter les souches en elles-mêmes. Pour la résistance, une catégorie est appliquée en fonction des marqueurs de résistance détectés sur la souche, selon le tableau 2.2.

Ces annotations portent sur les caractéristiques génomiques, mais les études contiennent aussi parfois des caractéristiques phénotypiques. Lors de l'indexation il est possible d'enrichir les souches à l'aide d'annotations phénotypiques issues de divers articles. Enfin, le NCBI contient un grand nombre de métadonnées sur chaque échantillon, par exemple le pays de prélèvement, le sexe de l'hôte, le type de tuberculose, l'étude associée, les auteurs, et parfois des caractéristiques phénotypiques. Ces informations sont automatiquement récupérées et indexées avec chaque souche.

Les métadonnées fournies par le NCBI ne sont pas normalisées, l'extraction repose donc sur un ensemble d'heuristiques afin de normaliser les données. Pour déterminer ces heuristiques, l'intégralité des champs de métadonnées du NCBI a été extrait. Chaque champ a été manuellement catégorisé en fonction de la donnée qu'il contient (tableau B.3). La majorité de la normalisation est effectuée avec une liste de mots clés (tableau B.5, tableau B.4). L'extraction de la localisation géographique (tableau B.2) requiert des étapes supplémentaires. Quand la localisation est donnée sous forme de coordonnées, un algorithme a été développé pour effectuer un parsing des coordonnées avant d'effectuer un géocodage inversé.

Toutes ces annotations sont mises en cache, au vu de la masse de données à traiter, afin de ne pas surcharger les sources de données publiques.

3.3.3 Indexation

Avec le volume massif de données, plus de 125 000 souches, 2 500 000 variants et un total de plus de 100 000 autres marqueurs (RD, IS, etc...), cela représente près de 200 000 000 de liens entre l'ensemble des données. Il est nécessaire de pouvoir effectuer des requêtes booléennes, des statistiques, des calculs plus complexes comme des comparaisons pour connaître les marqueurs en commun.

De nombreuses bases de données seraient capables d'effectuer ces opérations, cependant il est souhaité d'effectuer ces calculs avec des temps de réponse très faibles (inférieurs à quelques seconde). En effet, lors de l'utilisation de la plateforme, de nombreuses opérations peuvent être effectuées par l'utilisateur selon un grand nombre de critères, un temps de réponse supérieur réduirait ainsi grandement la productivité. Les bases de données avec index inversé, comme utilisées dans les moteurs de recherche (Google, etc...) démontrent une capacité à traiter un important volume de données avec un temps de réponse très court en appliquant des opérateurs booléens complexes. Ces bases de données sont initialement utilisées pour rechercher du texte dans un large corpus, mais on peut remarquer que ces opérations de recherche textuelles sont peu différentes des opérations que l'on souhaite effectuer sur les données génomiques. Ainsi en représentant toutes les données génomiques dans une forme adaptée à ces index inversés, il est possible d'exploiter leur efficience pour notre cas d'usage.

Un pipeline d'indexation, développé en C# utilisant l'Application Programming Interface (API) Dataflow, est utilisé pour lire les rapports JSON du pipeline génomique et pour indexer les données dans Elasticsearch 8. Le rapport est d'abord lu en mémoire, divisé en 4 types de données (souche, variant, RD, IS), annoté et enfin indexé dans des indices séparés. La performance d'indexation est proche de 20 rapports par seconde en

utilisant un processeur AMD Ryzen 7 2700 à 8 cœurs @ 3.20GHz avec 32GiB de RAM, avec une instance locale de la base de données.

3.3.4 Calcul de l'exclusivité

Au-delà de pouvoir générer des statistiques à partir de n'importe quel ensemble de souches sélectionné, la plateforme permet de réaliser des calculs avancés, notamment d'exclusivité.

L'exclusivité des caractéristiques permet de sélectionner un sous-ensemble d'isolats cliniques (appelé ensemble de premier plan) parmi l'ensemble (appelé ensemble de fond) de toutes les séquences génomiques de la base de données, et de déterminer si les caractéristiques de ces isolats (variants, RD, IS) sont exclusives au sous-ensemble. Une caractéristique est exclusive à 100% si elle est présente dans toutes les souches de l'ensemble de premier plan et dans aucune des souches de l'ensemble de second plan. TB-Annotator est capable d'exécuter une telle analyse sur une très grande quantité de souches en tirant parti des ordinaux globaux de l'index inversé.

À l'aide d'un script personnalisé, le coefficient de Jaccard ($J(A,B) = \frac{|A \cap B|}{|A \cup B|}$) est calculé pour être utilisé comme score pour chaque caractéristique des souches, A étant l'ensemble des souches sélectionnées et B l'ensemble des souches ayant la caractéristique. Pour chaque caractéristique (variants, RD, IS, ...), on calcule la fréquence d'apparition de la caractéristique dans l'ensemble de premier plan et dans celui de second plan. Puis on calcule la taille des deux ensembles. Ces valeurs permettent de calculer ensuite le coefficient selon la formule suivante :

$$\text{exclusivity}(A,B) = J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{f_{sub}}{f_{sup} - f_{sub} + s_{sub}}$$

Où :
- A est l'ensemble des souches sélectionnées,
- B est l'ensemble des souches ayant la caractéristique,
- f_{sub} est la fréquence de la caractéristique dans l'ensemble de premier plan,
- f_{sup} est la fréquence de la caractéristique dans l'ensemble de fond,
- s_{sub} est la taille de l'ensemble de premier plan,
- $sub \subseteq sup$.

Cette fonctionnalité est particulièrement utile pour trouver les meilleurs marqueurs d'un sous-ensemble de souches donné. Par exemple, en sélectionnant les souches d'une branche d'un arbre phylogénétique, il est possible de trouver les SNP exclusifs à ces souches.

On conserve par ailleurs le nombre de souches ayant la caractéristique dans l'ensemble de premier plan, le nombre de souches manquant la caractéristique, le nombre de souches hors de l'ensemble de premier plan ayant la caractéristique. Quand le score d'exclusivité est peu informatif, ces informations permettent d'aider à trouver un ensemble de premier plan partageant un nombre supérieur de caractéristiques.

3.3.5 Calcul de la matrice des distances

Un autre calcul, utilisé notamment pour déterminer si un ensemble de souches est un cluster de transmission ou non, est celui de la distance entre les souches. La distance entre deux souches étant définie comme le nombre de SNPs n'étant pas en commun entre deux souches. On estime par ailleurs le taux d'acquisition de SNPs à 0.5 par an [294] et qu'avec une distance inférieur à 5 SNPs les souches sont épidémiologiquement liées.

Pour effectuer ce calcul, l'utilisateur sélectionne un ensemble de souches, la base de données est lue en streaming pour extraire toutes les caractéristiques de cet ensemble. Les caractéristiques utilisées sont personnalisables (SNP, SNP dans des régions polymorphiques et indels). Pour chaque paire de souches de l'ensemble on calcule la distance, puis

on construit la matrice des distances, qui est donc une matrice symétrique. On détermine ensuite la moyenne des distances ainsi que l'écart type. Toutes ces données, incluant la matrice, sont renvoyées à l'utilisateur.

$$D = [d_{ij}]$$

$$d_{ij} = |S_i \cup S_j| - |S_i \cap S_j|$$

Où :
— D est la matrice des distances,
— S_i et S_j désignent respectivement l'ensemble de SNPs des souches i et j.

De par la méthode utilisée, la matrice des distances est seulement calculable pour de petits ensembles de souches.

3.3.6 Recherche de souches similaires

La recherche de souches similaires permet, à partir d'une souche donnée, de trouver d'autres souches partageant des caractéristiques (SNPs, etc...) proches. Pour cela il ne suffit pas de rechercher les souches avec le plus grand nombre de caractéristiques en commun car beaucoup de caractéristiques sont communes à de nombreuses souches, la majorité des caractéristiques sont donc peu informatives. Par exemple, les souches de L7 sont très peu représentées dans la base de données (< 50 souches). Ces souches partagent de nombreuses caractéristiques avec toutes les autres souches, mais elles partagent entre elles des caractéristiques uniques et exclusives. Et c'est cette dernière catégorie qui est la plus informative.

Ce problème est bien connu dans le domaine de l'IR (information retrieval). Il s'agit de rechercher des documents similaires à un document, en se basant sur la fréquence des mots. Une de ces méthodes est basée sur term frequency–inverse document frequency (tf-idf), qui permet de mesurer l'importance d'un mot dans un document, en prenant en compte le fait que certains mots ont une fréquence plus élevée que d'autres. Les mots d'un document ayants le score tf-idf le plus élevé sont les meilleurs représentants de ce document. Nous pouvons appliquer ce même principe sur les caractéristiques génomiques des souches. Les caractéristiques ayant le tf-idf le plus élevé sont les meilleures représentantes de la souche. Il est ensuite possible de chercher les souches ayant ces caractéristiques.

Le tf-idf est le produit de deux statistiques, la fréquence des termes et la fréquence inverse des documents. Nous utiliserons dans la suite «caractéristiques» et «souche». La fréquence des caractéristiques est définie comme :

$$\mathrm{cf}(c, s) = \frac{f_{c,s}}{\sum_{c' \in s} f_{c',s}}$$

Où :
— c est la caractéristique,
— s est la souche,
— $f_{c,s}$ est le nombre de fois que la caractéristique est présente dans la souche,
— le dénominateur correspond au nombre total de caractéristiques.

De par la nature de certaines données, comme les SNPs, le numérateur sera systématiquement égal à 1 et le dénominateur sensiblement égal pour toutes les souches, cf a donc peu d'influence en l'état sur le résultat final. La fréquence inverse des souches est définie comme :

$$\mathrm{isf}(c, S) = \log \frac{N}{|\{s : s \in S \text{ and } c \in s\}|}$$

Où :

— $N = |S|$ est le nombre total de souches,
— $|\{s : s \in S \text{ and } c \in s\}|$ est le nombre de souches ayant la caractéristique t.

Enfin le score final est défini comme :

$$\text{score}(c, s, S) = \text{cf}(c, s) \cdot \text{isf}(c, S)$$

Un score élevé est atteint par une fréquence de la caractéristique dans la souche et une faible fréquence de la caractéristique dans l'ensemble des souches.

Lors de la recherche de souches similaires, on commence par extraire les caractéristiques avec le score le plus élevé. Ensuite, on crée une requête conjonctive avec ces caractéristiques pour obtenir une liste de souches similaire. Ainsi, en cherchant les souches similaires à une L7, les SNPs exclusifs de la L7 sont relativement rares et lors de la recherche ces SNPs les autres L7 seront surreprésentées.

3.3.7 Arbres phylogénétiques

Dans le cas où l'on souhaite réaliser une étude phylogénétique sur un ensemble de souches, le processus est long et fastidieux. Il est nécessaire de lancer un pipeline pour extraire les marqueurs des souches, créer une matrice de ces marqueurs, créer l'arbre phylogénétique. Et ensuite il est souvent impossible de comparer cet arbre avec l'ensemble de la diversité du MTBC.

Une des fonctionnalités les plus avancées de la plateforme est la possibilité d'intégrer des arbres phylogénétiques personnalisés. Pour cela la matrice des caractéristiques peut être extraite directement de la base de données, l'arbre calculé par l'utilisateur à l'aide d'outils comme RAxML-ng [259], et l'arbre analysé ensuite à nouveau en utilisant toutes les fonctionnalités de la plateforme.

Du point de vue de l'utilisateur, on commence par la sélection d'un ensemble de souches à exporter en utilisant les filtres de la plateforme. Puis l'utilisateur sélectionne les différentes caractéristiques à exporter (SNPs, ISs, etc...). Ensuite la plateforme propose le téléchargement de l'alignement au format FASTA. En interne, la création d'alignements se fait en interrogeant la base de données par batch et en calculant dynamiquement une matrice entre chaque document et les caractéristiques qu'il contient. Cette matrice est ensuite utilisée pour générer un alignement au format FASTA. Cela permet de conserver les nucléotides dans l'alignement et d'utiliser les différents modèles d'évolution qu'offrent les outils d'inférence d'arbres phylogénétiques. Dans le cas où d'autres caractéristiques que les SNPs sont exportées, un alignement binaire est exporté par la plateforme, car seuls les SNPs peuvent être représentés sous forme de matrice. L'utilisateur devra donc configurer l'outil d'inférence d'arbres pour prendre en entrée une matrice binaire. Les détails des colonnes sont également exportés à la fin du fichier FASTA dans des commentaires FASTA. Cela permet d'obtenir la correspondance entre la colonne de la matrice et la caractéristique associée. Cette fonctionnalité d'export sert donc à la fois de base pour la réalisation d'arbres phylogénétiques mais aussi de toute autre forme d'analyse externe.

Après l'utilisation de RAxML-ng, l'utilisateur obtient un fichier Newick, qui est la représentation de l'arbre phylogénétique inféré. Il est alors possible de déposer le fichier Newick sur la plateforme pour permettre d'analyser l'arbre avec toutes les fonctionnalités de calcul statistique et de filtrage de TB-Annotator. Pour cette fonctionnalité de graphe de souches, un algorithme effectue le parsing du fichier Newick dans le navigateur. On obtient ainsi une représentation de l'arbre en mémoire. Cette représentation est très légère car elle ne contient que les numéros d'identification des souches et la structure de l'arbre. L'étape suivante consiste à dessiner l'arbre dans le navigateur, la tâche est difficile au vu de la masse de données à afficher, un arbre pouvant contenir plusieurs milliers de souches et autant de connexion à afficher entre les souches.

Il est nécessaire, à partir de l'arbre, de déterminer la position de chaque sommet et arête sur un plan (dimension 2). Pour cela on utilise un algorithme de layout (rendu/place-

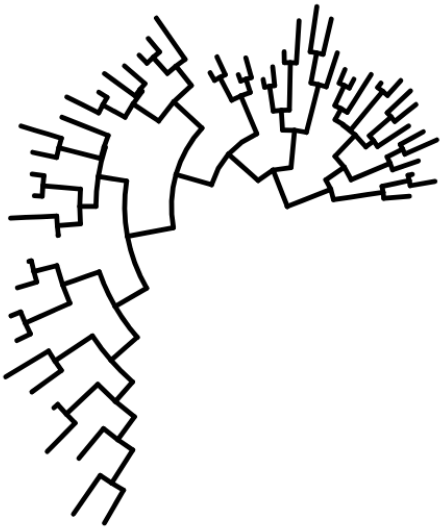


FIGURE 3.10 – Arbre dessiné selon l'algorithme de layout à angle égal [83]

ment). Il en existe plusieurs avec des caractéristiques et des représentations différentes. Par exemple, les layout circulaires représentent l'arbre avec les sommets disposés sous forme de cercle. L'objectif de la plateforme étant de permettre des analyses à grande échelle, et le dessin s'effectuant dans le navigateur, il a été choisi un algorithme de faible complexité. L'arbre est rendu en utilisant un algorithme de layout à angle égal (figure 3.10) [83]. En partant de la racine de l'arbre, on divise un cercle (360 degrés) par le nombre d'enfants proportionnellement au nombre de descendants. Les enfants seront placés dans cet angle selon la longueur de branche. On continue récursivement pour les nœuds enfants en divisant à nouveau l'angle qui leur a été attribué. Le dessin s'effectue en utilisant la technologie WebGL et la carte graphique de l'utilisateur.

Afin de pouvoir utiliser les fonctionnalités de la plateforme, il est nécessaire de faire le lien entre l'arbre mis en ligne par l'utilisateur et les données de cette dernière. Pour cela il est imposé que les labels de l'arbre correspondent aux identifiants des souches associées. Lors d'une sélection, le nom de chaque nœud dans le graphe est extrait pour être envoyé au serveur et utilisé comme liste d'accession. Cela permet l'utilisation du calcul d'exclusivité des caractéristiques avec des arbres phylogénétiques arbitraires. La liste de toutes les souches affichées est utilisée comme ensemble de fond pour la comparaison d'exclusivité (l'ensemble de la base de données peut également être utilisé comme ensemble de fond).

Un exemple d'utilisation de ces fonctionnalités est la recherche de marqueurs phylogénétiques. Après calcul d'un arbre, il est possible de rechercher visuellement des marqueurs exclusifs à des lignées. Cette recherche est faite en prenant en compte l'ensemble des souches publiquement disponibles et non uniquement l'arbre chargé. Ensuite, grâce à une fonctionnalité de stockage des filtres, il est possible de sauvegarder la définition de la lignée avec le marqueur exclusif trouvé. Cela permet par exemple de créer des ensembles de marqueurs pour un barcoding rapide des souches.

3.3.8 Utilisation via API

De par la flexibilité, la plateforme d'analyse couvre un grand nombre de besoins, aussi bien en termes de filtrage qu'en termes de calcul statistique. Pour les utilisateurs avancés, ces fonctionnalités restent parfois insuffisantes. Nous souhaitons néanmoins permettre l'exploitation de cette masse de données génomiques ainsi que l'index quand l'opération n'est pas gérée nativement par la plateforme. C'est pour cela que TB-Annotator permet

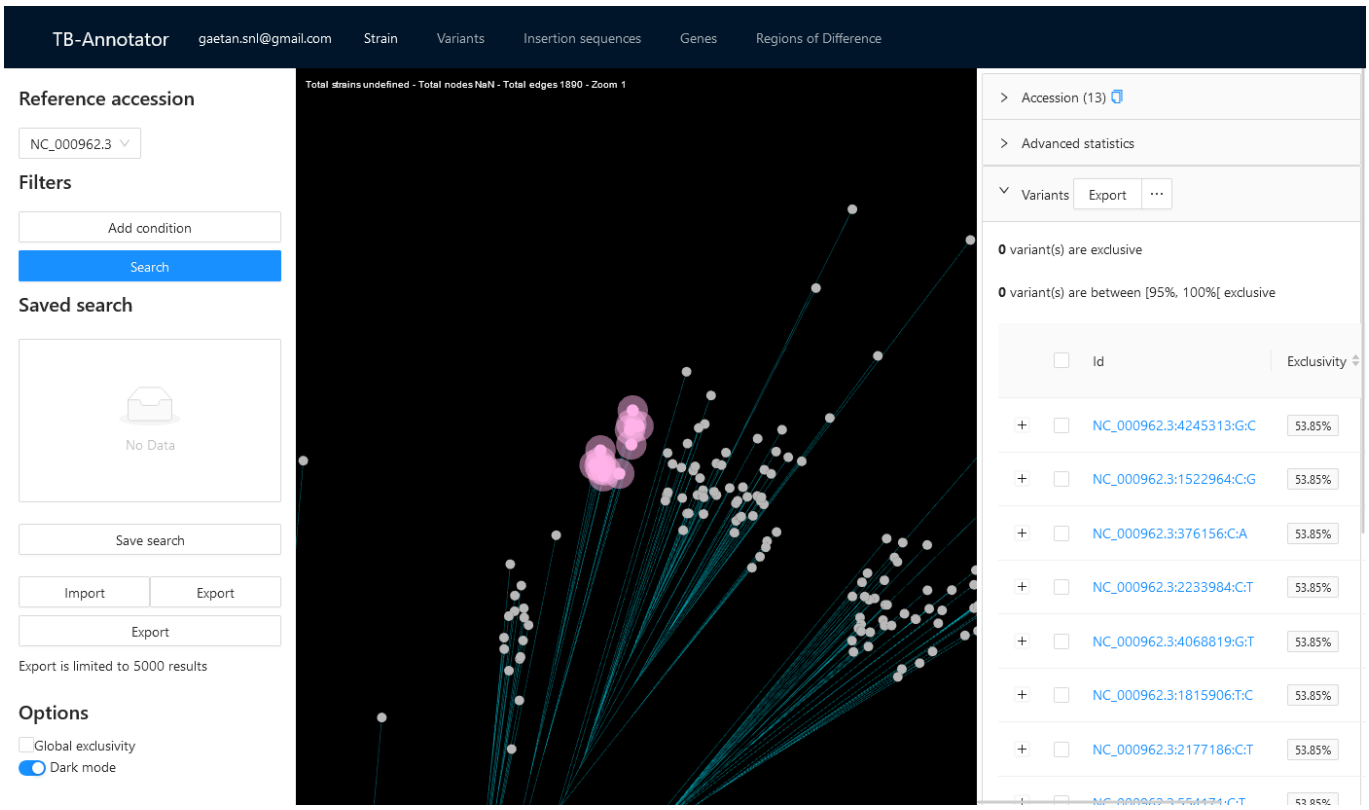


FIGURE 3.11 – Vue d'un arbre phylogénétique dans la plateforme TB-Annotator. 13 souches sont sélectionnées graphiquement par l'utilisateur sur l'arbre et l'exclusivité des variants est calculée par rapport aux souches du reste de l'arbre.

aussi un accès via API.

3.3.9 Comparaison avec un arbre phylogénétique global

Dans une première tentative, un serveur d'analyse basé sur un arbre phylogénétique unique a été développé, l'idée étant de calculer un arbre phylogénétique global représentatif de la diversité du MTBC. Cet arbre sert alors de point de départ à différents calculs. Par exemple il est possible de sélectionner les souches issues d'une branche de l'arbre, et lister les caractéristiques phylogénétiques communes à cette branche. Ce serveur dispose d'une interface web et ne requiert pas d'installation, ce qui permet de s'adresser au plus grand nombre, sans nécessiter de compétence technique, la difficulté d'exploitation de ce type d'outil étant une barrière majeure à leur utilisation.

Cette approche se heurte rapidement à une limitation majeure : le temps de calcul requis pour un arbre précis d'une telle taille empêche toute possibilité d'exploiter plus d'une dizaine de milliers de génomes. Rien qu'à cette échelle réduite, il est nécessaire de développer du code hautement optimisé pour permettre un affichage dans un navigateur web. Il a donc été décidé de réaliser cet arbre sur un set réduit de 16000 génomes.

L'application web de cette première version est hautement optimisée afin de gérer l'affichage et l'analyse de ces 16000 génomes. L'arbre phylogénétique est calculé hors ligne mais l'intégralité des autres calculs est effectuée en temps réel sur l'interface. La communication est réalisée par protocole binaire (GRPC) et le rendu de l'arbre est effectué avec WebGL. La base de données est développée en C# sous forme d'un index inversé dont la persistance est effectuée à l'aide de RocksDB.

Ce système présente plusieurs limitations significatives qui entravent son efficacité. Premièrement, le fait que le sous-ensemble des souches soit limité par définition implique que l'analyse ne peut pas couvrir l'ensemble des données collectées du MTBC. En outre, les performances ne permettent pas d'augmenter la taille de ce sous-ensemble de souches,

limitant ainsi l'étendue des analyses possibles. Cette contrainte est aggravée par l'incapacité d'analyser plus d'un jeu de données à la fois, interdisant par exemple la visualisation distincte des souches appartenant à une lignée spécifique. La mise à jour de l'arbre phylogénétique est fastidieuse, nécessitant l'intervention manuelle d'une personne. D'autre part, le système ne permet pas la recherche de données autres que des souches, comme les SNP, ce qui limite l'amplitude des recherches possibles. Le calcul de l'exclusivité est particulièrement complexe et ne pourra jamais absorber la quantité de données requises pour analyser l'ensemble du NCBI, rendant ce type d'analyse impraticable. Enfin, la réalisation d'analyses statistiques est loin d'être aisée, complexifiant ainsi l'exploitation des données pour des études approfondies. Cette version a néanmoins permis de réaliser plusieurs découvertes exposées dans les chapitres suivants.

3.4 Comparaison avec d'autres outils

Malgré le manque de standardisation des pipelines d'analyse WGS du MTBC et le besoin de bases de données pour centraliser les données WGS à un niveau mondial [180], de nombreux outils existent actuellement. Cette section propose une présentation de différents outils ainsi que leur limitations par rapport à une plateforme entièrement intégrée et possédant des données homogènes, comme c'est le cas pour TB-Annotator.

3.4.1 Outils d'analyse WGS

TBprofiler [42] est un autre outil d'analyse de données NGS pour *Mycobacterium tuberculosis* qui se concentre sur la détection des mutations de résistance aux médicaments. TBprofiler propose également une interface web, mais son champ d'application est limité à l'analyse de la résistance aux médicaments. TB-Annotator va au-delà en permettant l'analyse et la visualisation des insertions, des régions de différence, et des gènes manquants, fournissant ainsi une analyse plus complète des génomes.

CASTB (Comprehensive Analysis Server for the *Mycobacterium tuberculosis* Complex [129]) est une base de données et un serveur web qui permet l'analyse des séquences du génome complet des souches du complexe *M. tuberculosis*. L'outil fournit des informations sur la résistance aux médicaments et l'identification des lignées à travers une série d'analyses, telles que la phylogénie basée sur les SNP et la prédiction des mutations de résistance aux médicaments. Cependant, contrairement à TB-Annotator, CASTB n'est pas conçu pour être utilisé pour l'analyse en temps réel de données génomiques à grande échelle, et il ne dispose pas d'un pipeline d'annotation complet pour détecter les variations génomiques au-delà des mutations de résistance connues.

Mykrobe [22] est une suite d'outils pour l'identification de la résistance aux médicaments chez les pathogènes bactériens, y compris *M. tuberculosis*. L'outil fournit des prédictions pour la résistance aux médicaments de première et deuxième ligne, et il permet également l'identification des lignées pour les isolats. Cependant, Mykrobe n'est pas conçu pour l'annotation complète des variations génomiques au-delà des mutations de résistance.

Comparé à ces outils, TB-Annotator fournit un pipeline complet pour la détection des variations génomiques, y compris les SNPs, les insertions, les délétions, et les réarrangements génomiques à grande échelle. De plus, TB-Annotator permet l'annotation des variations génomiques au-delà des mutations de résistance connues, permettant aux chercheurs d'étudier l'évolution et la transmission des souches de *M. tuberculosis* de manière plus détaillée. En outre, TB-Annotator propose une plateforme web conviviale qui permet l'analyse et la visualisation en temps réel des données génomiques à grande échelle. Dans l'ensemble, TB-Annotator se démarque comme un outil unique offrant une approche complète pour l'annotation et l'analyse des données génomiques dans le contexte de la recherche sur *M. tuberculosis*.

3.4.2 Profilage génétique

MIRU-VNTRplus [7] est un autre outil d'analyse en ligne conçu pour l'identification, le typage et la comparaison phylogénétique des souches de MTBC. Cet outil est basé sur la méthode MIRU-VNTR [269]. MIRU-VNTRplus offre une plateforme pour comparer de nouveaux isolats à une base de données de référence, permettant aux chercheurs et aux cliniciens de mieux comprendre la diversité génétique, l'épidémiologie et les modes de transmission de la tuberculose. Cet outil précède le développement du NGS et est donc très différent du nôtre : il se concentre sur un seul sujet d'analyse (les MIRU-VNTR), d'origine exclusivement expérimentale pour l'instant car il reste très difficile de déduire le nombre de copies des MIRU-VNTR à partir des SRAs [131], et il traite un seul isolat à la fois (celui soumis par l'utilisateur). Le MIRU-VNTR, bien qu'ayant été considéré pendant presque une décennie comme une norme, a montré en 2011 qu'il était moins discriminant que le WGS pour la surveillance génomique de la transmission de la tuberculose [96], et est également moins efficace que le WGS pour détecter les infections mixtes [35]. Pour l'instant, seuls les longs fragments permettent de reconstruire le locus MIRU-VNTR, grâce soit à MIRU-reader soit à MIRU-profiler [220, 275].

Une limitation est que TB-Annotator n'est pas capable de traiter les données d'extraction expérimentales de MIRU-VNTR, mais les fichiers de séquences. Il n'est actuellement pas en mesure d'analyser ces MIRU-VNTRs, car la taille des séquences ne permet pas de compter les copies des motifs d'intérêt. Cependant, il permet d'extraire les sites d'insertion IS*6110* à partir des SRAs, permettant ainsi de fournir une résolution élevée pour la surveillance épidémiologique, supérieure à ce qui peut être obtenu en utilisant les données de MIRU-VNTR [247].

Parmi les pipelines récemment développés pour analyser les données WGS sur MTBC, nous devrions également mentionner SAM-TB et SimpiTB [308, 66].

3.4.3 Annotations de gènes et profils d'expression

TBDB [53] est une base de données contenant des informations sur les gènes, les protéines et les mutations de résistance aux médicaments de *M. tuberculosis*. Elle sert de ressource pour les chercheurs en consolidant des données provenant de diverses sources et en fournissant des outils pour l'analyse et la visualisation des données. Bien que TBDB offre une riche collection d'informations génomiques, elle ne propose pas de pipeline d'analyse pour les données de séquençage NGS, contrairement à TB-Annotator. TB-Annotator analyse et annote non seulement les données génomiques, mais fournit également des capacités de recherche avancées et une plateforme pour effectuer des analyses sur les données obtenues.

TubercuList [54] et MycoBrowser [141] sont des bases de données en ligne contenant des informations génomiques annotées pour *Mycobacterium tuberculosis*. Ces ressources permettent l'exploration des gènes et des protéines, fournissant des informations détaillées sur la fonction des gènes, les ontologies des gènes et les voies métaboliques. Cependant, elles ne proposent pas les capacités d'analyse, de visualisation et de recherche qui sont au cœur de TB-Annotator. L'approche complète de TB-Annotator pour l'analyse des données de séquençage NGS, ainsi que ses fonctionnalités avancées de recherche et de visualisation, le distinguent de ces bases de données.

PATRIC [297] est une ressource en ligne pour les pathogènes bactériens, offrant des outils pour l'analyse génomique et comparative. PATRIC couvre une vaste gamme d'espèces bactériennes, y compris *M. tuberculosis*. Bien qu'il fournisse des données et des outils d'analyse précieux pour les chercheurs travaillant sur divers pathogènes bactériens, il n'est pas spécifiquement adapté à *M. tuberculosis* et ne propose pas le même niveau de profondeur d'analyse, de visualisation et de données spécifiques aux souches que TB-Annotator. TB-Annotator est conçu exclusivement pour *M. tuberculosis*, permettant une analyse plus ciblée et complète de cet organisme particulier.

BioCyc [143] est une collection de bases de données de voies métaboliques et génomiques qui couvrent divers organismes, y compris *Mycobacterium tuberculosis*. BioCyc offre des renseignements précieux sur les voies métaboliques, la fonction des gènes et les annotations génomiques. Cependant, il ne fournit pas les capacités complètes d'analyse génomique, de visualisation et de recherche que propose TB-Annotator. Une fois encore, TB-Annotator est spécifiquement conçu pour *M. tuberculosis*, ce qui lui permet de fournir une vue plus détaillée et ciblée des caractéristiques génomiques de cet organisme, ainsi que des options avancées de recherche et d'analyse.

3.4.4 Autres pipelines et sites web spécialisés

Enterobase [51] est une base de données de pathogènes bactériens accessible au public, qui inclut également *Mycobacterium tuberculosis*. Elle permet aux utilisateurs de soumettre des données de séquençage et de les comparer avec d'autres isolats pour identifier des clusters d'épidémies et suivre la propagation de la résistance aux antimicrobiens [315]. Bien qu'elle offre une fonctionnalité similaire à celle de TB-Annotator, elle ne fournit pas le même niveau d'annotation et d'analyse des caractéristiques génomiques pour le MTBC. La vitesse de calcul est nettement inférieure et ne permet pas l'analyse en temps réel, l'analyse étant effectuée en arrière-plan. Les fonctionnalités de recherche des marqueurs sont aussi absentes. De plus, Enterobase n'inclut pas actuellement toutes les données de séquençage disponibles de *M. tuberculosis*, alors que TB-Annotator intègre une grande partie des données SRA disponibles publiquement.

PathogenSeq [52] est une plateforme web qui fournit des outils pour l'analyse des données génomiques microbiennes, y compris *M. tuberculosis*. Elle permet aux utilisateurs de télécharger leurs propres données ou d'utiliser des ensembles de données préchargés et offre des outils pour l'assemblage, l'annotation et la génomique comparative. Cependant, contrairement à TB-Annotator, PathogenSeq se concentre davantage sur la fourniture d'outils d'analyse de données brutes plutôt que sur des données pré-annotées. De plus, elle n'inclut pas certaines des fonctionnalités de TB-Annotator, comme la possibilité de rechercher des caractéristiques génomiques spécifiques, d'examiner des arbres phylogénétiques et d'effectuer une analyse d'exclusivité des caractéristiques.

Nextstrain [112] est une plateforme web qui offre un suivi en temps réel des épidémies de maladies infectieuses mondiales, comprenant aussi *M. tuberculosis*. Elle permet aux utilisateurs de visualiser des arbres phylogénétiques de séquences virales et bactériennes, de suivre la propagation des épidémies et d'identifier de nouveaux variants génétiques. Bien qu'elle offre certaines fonctionnalités similaires à TB-Annotator, elle ne fournit cependant pas le même niveau d'annotation et d'analyse des caractéristiques génomiques, ni certaines des options de recherche et de filtrage de TB-Annotator. De plus, Nextstrain se concentre davantage sur le suivi en temps réel de la propagation des maladies, tandis que TB-Annotator se concentre sur la fourniture d'une base de données exhaustive des caractéristiques génomiques de *M. tuberculosis*.

CNGBdb [50], enfin, est une base de données génomiques accessible au public, comprenant notamment *M. tuberculosis*. Elle fournit des outils pour la recherche et l'analyse des données génomiques, ainsi que pour la visualisation et la comparaison des données. Elle ne fournit cependant pas le même niveau d'annotation et d'analyse des caractéristiques génomiques, ni certaines des options de recherche et de filtrage de TB-Annotator. De plus, CNGBdb n'inclut pas toutes les données de séquençage disponibles de *M. tuberculosis*, alors que TB-Annotator le fait.

Chapitre 4

Apports pour la taxonomie du MTBC

4.1 Connexion entre deux sites historiques d'épidémies de tuberculose au Japon, sur l'île de Honshu, par une nouvelle sous-lignée ancestrale L2 de *Mycobacterium tuberculosis*

En rassemblant 680 Sequence Read Archives publiques provenant d'isolats du MTBC, dont 190 appartiennent à la lignée 2 *Beijing*, et en utilisant TB-Annotator, qui analyse plus de 50 000 caractères (au moment de cette étude), nous décrivons ici une nouvelle sous-lignée L2 à partir de 20 isolats trouvés dans la province de Tochigi, (Japon), que nous désignons sous le nom de *asia ancestral 5* (AAnc5). Ces isolats présentent un certain nombre de critères spécifiques (42 SNPs) et leur distance intra-cluster suggère une transmission historique et non épidémiologique. Ces isolats présentent une mutation dans rpoC, et ne répondent à aucun des critères de la lignée *modern Beijing*, ni à ceux des autres lignées *ancestral Beijing* décrites jusqu'à présent. Les isolats de *asia ancestral 5* ne possèdent pas les caractéristiques mutT2 58 et ogt 12 de *modern Beijing*, mais possèdent des SNPs caractéristiques de *ancestral Beijing*. En consultant la littérature, nous avons trouvé un isolat de référence ID381, décrit à Kobe et Osaka appartenant au groupe «G3», partageant 36 des 42 SNPs spécifiques trouvés dans AAnc5. Nous avons également évalué la position intermédiaire de la sous-lignée *asia ancestral 4* (AAnc4) récemment décrite en Thaïlande et proposons une classification améliorée de L2 qui inclut désormais AAnc4 et AAnc5. En augmentant les souches dans TB-Annotator à environ 3000 génomes (dont 642 appartenant à L2), nous avons confirmé nos résultats et découvert des branches ancestrales historiques supplémentaires de L2 qui restent à étudier plus en détail. Nous présentons également, en outre, des données anthropologiques et historiques sur l'histoire de la tuberculose en Chine et au Japon, ainsi qu'en Corée, qui pourraient soutenir nos résultats sur l'évolution de L2. Cette étude montre que la reconstruction de l'histoire précoce de la tuberculose en Asie est susceptible de révéler des motifs complexes depuis son émergence.

4.1.1 Introduction

Avec 9,9 millions de nouveaux cas en 2019 et 500 000 cas de résistance à plusieurs médicaments, la tuberculose est loin d'être éradiquée. Parmi les 9 lignées reconnues (L1 à L9) décrites dans le MTBC, la lignée L2 suscite un grand intérêt [74, 3, 31, 133, 181]. De très grandes épidémies de L2 ont été montrées comme ayant émergées indépendamment dans le monde entier [189]. Bien que l'origine de L2 soit soupçonnée être en Chine et que L2 soit prédominante en Asie de l'Est, son lieu et son moment exact d'émergence sont encore très débattus [171, 165, 166]. Les isolats cliniques de L2 ont développé des

caractéristiques spécifiques de virulence et de résistance aux médicaments qui contribuent à leur succès épidémique [181, 47, 150]. D'un point de vue évolutif, L2 a développé un mode de vie avec un nombre élevé de copies de IS*6110*, ce qui pourrait avoir favorisé un phénotype hypermutateur qui aurait pu augmenter la virulence de certains de ces isolats [86, 111]. L'explosion épidémique de L2 a été détectée pour la première fois dans les années 90 et a été favorisée par des événements historiques et géopolitiques tels que (1) la chute de l'ex-URSS et les changements en Chine ; (2) un traitement de la TB mal individualisé et un suivi médical des prisonniers dans ces pays ; (3) l'augmentation de la part du commerce mondial de la Chine [76, 290]. La L2 a encouragé de nombreuses études pour comprendre l'émergence de la résistance aux médicaments et pour améliorer sa caractérisation génétique [31, 19, 188]. Cela a été réalisé progressivement et par la combinaison de l'analyse de marqueurs polymorphiques tels que IS*6110*-RFLP, MIRU-VNTR, RD, loci VNTR hypervariables, et enfin WGS [283, 132, 182, 295, 8, 179, 246].

La lignée L2 a été divisée en deux principales sous-lignées, L2.1 (*Proto-Beijing*) et L2.2 (*Beijing*), [246, 199, 88]. L2.1 a été décrite principalement dans le sud de la Chine, particulièrement dans la province du Guangxi et pourrait être aussi ancienne que 30 000 ans, ayant coévolué avec les populations d'Asie de l'Est [171]. Parmi ses caractéristiques, elle porte la délétion RD105 mais pas RD207 [258]. Des isolats rares de L2.1 ont été montrés comme étant devenus ultra-résistants aux médicaments [258]. L2.2 est définie par SIT1 ou des variantes, et est composée de plusieurs sous-lignées, ancestrales et modernes. L2.2.2 définit la sous-lignée *asian ancestral* 1 ; L2.2.1 regroupe toutes les autres [43]. RD181 est spécifiquement supprimée dans toutes les sous-lignées L2.2.1 [245]. Le passage de *ancestral Beijing* à *modern Beijing* est associé à la présence d'au moins une copie de IS*6110* dans la région dite NTF et à la présence de mutations dans les gènes de réplication-réparation-recombinaison (3R), parmi lesquelles la mutation mutT2 G->C en position 1286766 et la mutation ogt C->T au codon 12 position 1477596, par rapport à la séquence de référence MTBC $H_{37}Rv$ (`NC_000962.3`) [245, 310]. Jusqu'à récemment, *ancestral Beijing* incluait 3 lignées ancestrales asiatiques (*asia ancestral* 1, 2 et 3, AAnc1, AAnc2, AAnc3) [245], jusqu'à ce qu'un nouvel *asia ancestral* 4 (AAnc4), soit découvert dans le nord de la Thaïlande [4]. Les souches *modern Beijing* sont responsables de la plupart, mais pas de toutes, les récentes épidémies de MDR-TB [31, 111, 142].

Les grandes bases de données construites avec des données WGS permettent de développer une connaissance précise et complète de la diversité de L2 [31, 179, 30]. La génomique computationnelle permet désormais d'étudier en profondeur à la fois l'histoire globale et locale de L2 [181, 295, 179, 258, 140, 292, 209, 293, 185]. L'histoire évolutive des isolats de L2 est toujours débattue, tout comme leur datation précise d'émergence et leur origine géographique [179, 310]. Luo et al. ont suggéré que L2 pourrait être aussi ancienne que 30 000 ans [171]]. Merker et al. ont estimé le temps jusqu'au plus récent ancêtre commun (TMRCA) à environ 6100 et 5200 ans Before present (BP) pour les complexes clonaux définis par MIRU-VNTR BL7 et CC6 (les plus anciens) et environ 1500 ans BP pour CC5 (le plus récent) [179]. Liu et al. ont estimé la coalescence entre L2.1/*Proto-Beijing* et L2.2.2/*ancestral Beijing* à 2200 BP, et à 1300 ans BP pour la séparation entre toutes les lignées *ancestral Beijing* [166]. L'expansion de L2.1/*Proto-Beijing* aurait eu lieu 900 BP tandis que le *modern Beijing* serait apparu il y a seulement environ 500 ans [166]. Hirsch et al. ont suggéré que les populations humaines d'Asie de l'Est et des Philippines portant des lignées distinctes de MTBC pourraient n'avoir été séparées que de 240 à 1000 ans [121].

L'origine géographique de l'émergence de L2 est aussi floue que sa datation. Selon certains auteurs, le centre-nord et le nord-est de la Chine auraient pu être le centre initial de propagation [187]. Selon d'autres, basés sur des différences dans la prévalence des lignées ancestrales de L2 en Chine et sur la plus grande diversité génétique observée dans la province du sud-ouest de Guizhou, le sud de la Chine pourrait être le berceau de L2 [171, 310]. En effet, le Guizhou compte 17 minorités ethniques et la plupart des groupes

ethniques reconnus de Chine sont situés dans cette province. Plus généralement, l'Asie du Sud-Est montre une plus grande diversité génétique humaine que l'Asie du Nord-Est [126]. Plaidant en faveur d'une origine dans le sud de la Chine, la description récente d'un sous-lignage de L2 *asia ancestral* 4, a été faite dans le nord de la Thaïlande à Chiangrai, habitée depuis le 7ème siècle et peuplée par des minorités ethniques originaires du Sud de la Chine [4].

Des preuves de tuberculose sur des squelettes datant de l'âge du bronze ont été trouvées en Corée et au Japon [274, 273]. Au Japon, l'une des principales caractéristiques de l'histoire de la tuberculose, notamment chez les personnes âgées non vaccinées par le BCG, est la présence de souches ancestrales de L2 encore mal caractérisées [132, 179, 292, 185, 291]. La diversité MIRU-VNTR avait été montrée plus tôt comme étant assez importante dans les isolats de L2 provenant du Japon [187]. D'autres preuves basées sur MIRU-VNTR avaient suggéré que certaines souches ancestrales spécifiques de L2 pourraient être endémiques au Japon [311, 312, 197, 313]. Depuis la publication de ces études, le WGS de quelques isolats de référence a été publié à Kobe et Osaka, cependant ils n'ont pas été mentionnés dans la phylogénie simplifiée de L2 [246, 292].

Nous avons étudié un ensemble d'isolats de L2 de la préfecture de Honshu central au Japon, Tochigi [185]. Les 169 isolats cliniques de *M. tuberculosis* que nous avons étudiés provenaient de patients atteints de TB diagnostiqués en 2007 et en 2013 [185]. Les données WGS sur ces isolats ont été publiées après analyse en utilisant un pipeline bioinformatique, «CAST» par Iwai et al. [130]. Puisqu'une analyse génomique comparative de ces isolats avec d'autres L2 ancestraux n'avait pas été réalisée, nous avons inclus ces génomes dans notre base de données. Nous avons également inclus des SRAs étiquetés comme a*asia ancestral* 4 [4]. En utilisant notre pipeline, nous avons décrit la caractérisation génétique d'une nouvelle sous-lignée de L2 du Japon, nommée *asia ancestral 5* (AAnc5), qui semble être exclusive au Japon pour le moment. Nous avons également évalué les caractéristiques de la sous-lignée AAnc4 décrite en Thaïlande et fourni un cadre évolutif mondial mis à jour de la lignée L2 [245].

4.1.2 Sélection des génomes et méthode

Pour toutes les analyses, TB-Annotator a été utilisé pour extraire l'intégralité des marqueurs : recherche de SNPs selon des catalogues de référence, recherche de SNPs supplémentaires dans chaque isolat basée sur la séquence de référence $H_{37}Rv$, recherche de la présence/absence de gènes $H_{37}Rv$ tels qu'annotés dans mycobrowser, recherche de la présence/absence de régions de délétion, identification des sites d'insertion de toutes les séquences d'insertion connues dans *Mycobacterium tuberculosis* complex. Le locus CRISPR est reconstruit semi-automatiquement en utilisant un script dédié et précédemment publié (format 43/68/360 *spacers*) avec une attribution d'un tag Spoligo-International-Type (SIT) ; l'application produit une liste ordonnée de *spacers*/*repeat* avec variants et séquences d'insertion IS*6110* si présentes [106, 64, 225].

Génomes sélectionnés

Nous avons téléchargé un ensemble de 680 SRAs ; ces échantillons ont été sélectionnés pour représenter la diversité génomique de la TB (L1 à L9) décrite jusqu'à présent, y compris dans des articles de référence récents [199, 88]. La liste de ces SRAs est présentée dans le Tableau Supplémentaire S2 (liste de 680 SRA incluant 190 SRA L2) ; la base de données a été construite pour représenter toutes les sous-lignées L2 à l'exception de la sous-lignée Pacific RD150. D'après [4], nous avons sélectionné 28 SRAs étiquetés AAnc4. De [185], nous avions initialement inclus 158 SRAs, cependant 57 SRAs pour lesquels la couverture était soit trop faible, soit pour lesquels il était impossible de reconstruire le spoligotype en utilisant CRISPR-builder-TB ont été écartés [106, 225].

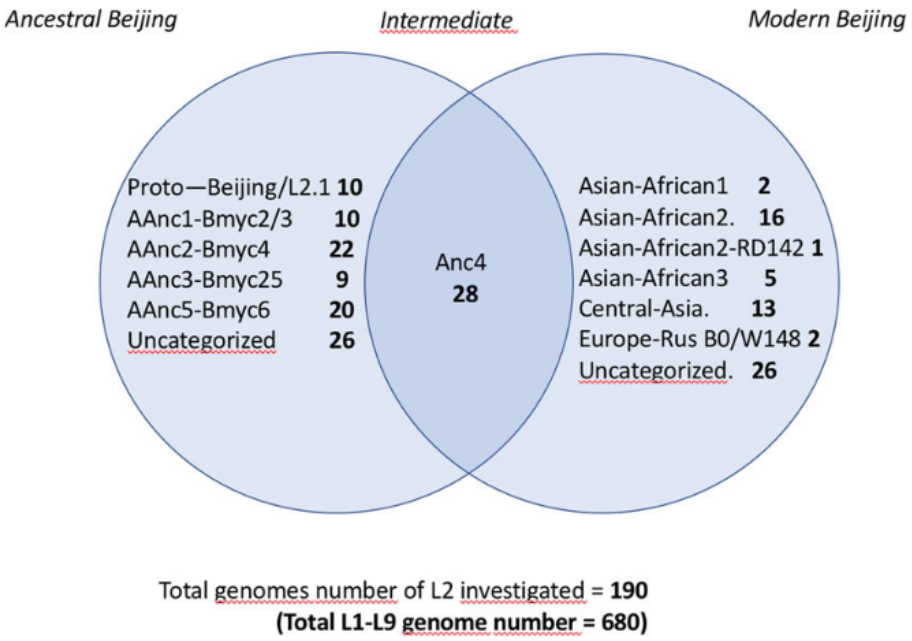


FIGURE 4.1 – Diagramme de Venn montrant la classification des 190 L2 SRA étudiés

Bioinformatique et méthodes phylogénétiques

Les scripts finaux permettent de produire une phylogénie basée sur la liste des caractères étudiés et utilisant RAxML et SplitsTree [260, 149]. Tous les calculs ont été effectués sur les installations du superordinateur «Mésocentre de Franche-Comté» (141 nœuds, 2292 cœurs, 9,27 To de mémoire, 74,2 TFlops de puissance CPU, 66,4 TFlops GPU), en utilisant les commandes adéquates. En dehors des résultats présentés dans le Tableau Supplémentaire S3, un arbre phylogénétique final est affiché graphiquement et des scripts python propriétaires permettent de réaliser des requêtes interactives et d'afficher les résultats [149]. La version actuelle de TB-Annotator à ce moment là incluait 6009 génomes et a confirmé nos résultats.

Classification basée sur les SNPs des sous-lignées L2

Afin d'attribuer les 680 SRAs sélectionnés aux lignées et sous-lignées connues, nous avons utilisé la liste de référence des marqueurs définie dans le Tableau Supplémentaire S1. Un diagramme de Venn simplifié montre la classification des 190 SRAs L2 en 125 isolats ancestraux et 65 isolats modernes (figure 4.1). Les résultats de la classification basée sur les SNPs pour les 190 SRA des isolats L2, tels que produits par TB-Annotator, se trouvent dans le Tableau Supplémentaire S3.

Méthodes et jeu de données utilisés pour définir une nouvelle sous-lignée ancestrale asiatique 5 (AAnc5)

Après l'évaluation des sous-lignées connues du MTBC (figure 4.2 a et b), la branche inconnue (figure 4.2 c) a été étudiée en détail : reconstruction *in silico* du locus CRISPR en utilisant CRISPR-builder TB [106, 225], analyse *in silico* MIRU-VNTR avec CAST[130] ; évaluation des SNPs et distance intra-cluster avec TB-Annotator ; il a été démontré que AAnc5 provenait uniquement des Sequence Read Archives d'isolats du Japon (Bioproject PRJDB3875). La pipeline bioinformatique, basée sur son interface graphique utilisateur, permet de sélectionner et d'afficher de nouveaux SNPs exclusifs ou partagés et des caractéristiques génétiques spécifiques, qui ont été ensuite approfondies. Les marqueurs génomiques précédents extraits de Wada et al. 2012 ont été comparés à nos résultats et se

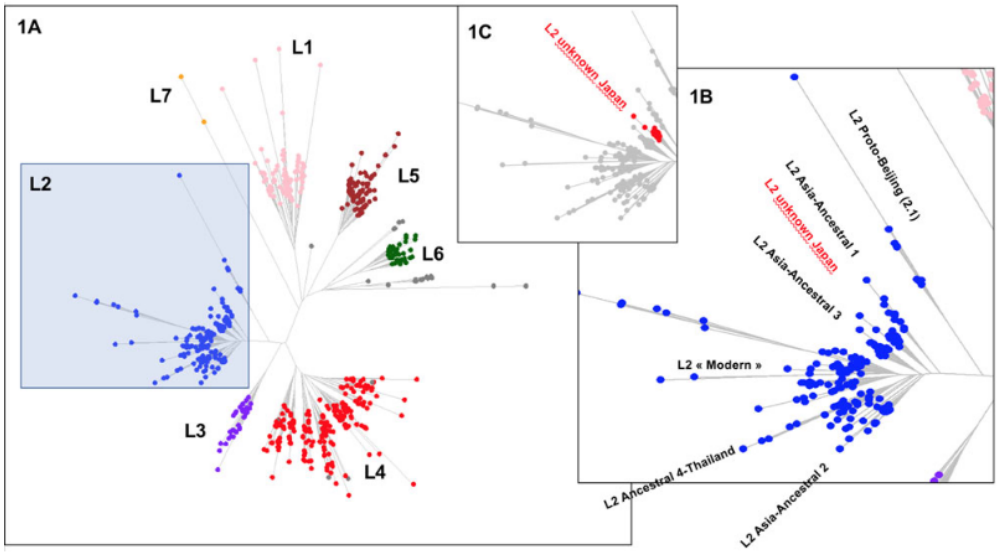


FIGURE 4.2 – Partie gauche (a) ; TB-Annotator arbre phylogénétique non enraciné sur 680 données dérivées de Sequence Read Archives. Les échantillons L2 sont montrés en bleu. Partie droite (b) ; zoom sur la Lignée 2 avec toutes les branches connues nommées sauf en rouge la nouvelle sous-lignée ancestrale du Japon inconnue que nous avons désignée comme asia ancestral 5. Partie centrale (c) ; focus sur la lignée inconnue du Japon.

trouvent dans le Tableau Supplémentaire S4 [292].

4.1.3 Découverte et caractérisation d'une nouvelle sous-lignée de L2

Découverte d'une nouvelle sous-lignée L2

Nous avons mis en œuvre un ensemble représentatif de SRAs de L2 comprenant 101 extraits de l'étude de 169 échantillons de la province de Tochigi, sur la plateforme TB-Annotator [185]. La classification des 190 génomes L2 étudiés est présentée dans un diagramme de Venn simplifié (figure 4.1) et l'arbre phylogénétique global produit est montré sur la Figure 2a–c. Tous les échantillons portaient le SNP définissant L2 (G497491A) et se trouvaient sur la même branche phylogénétique (figure 4.2 a). Sur la base de la recherche de SNP, et comme illustré sur la Figure 2b et c et plus en détail dans le Tableau Supplémentaire S3, parmi les *ancestral Beijing* (n = 125), nous avons trouvé *Proto-Beijing* (L2.1 ; n = 10), *asia ancestral* 1 (L2.2.2-AAnc1, n = 10), asia-ancestral 2 (AAnc2, n = 22), *asia ancestral*l 3 (AAnc3, n = 9), *asia ancestral* 4 (AAnc4, n = 28), et une branche inconnue (suggérée comme *asia ancestral* 5 ou AAnc5, n = 20) ; il restait 26 *ancestral Beijing* non classifiés. Figure 1 et Tableau Supplémentaire S3.

Quatre-vingt-treize isolats (y compris AAnc4), faisaient partie de L2.2.1 et portaient tous un mutT2 G1286766C qui définit traditionnellement un SNP caractéristique de *modern Beijing*. Parmi ces 93, cependant, seulement 65 échantillons (excluant AAnc4) portaient le second SNP signature moderne de L2, c'est-à-dire le C1477596T dans ogt, tandis que les 28 AAnc4 ne portaient pas ce SNP [245, 4]. Par conséquent, la désignation de L2 moderne devrait être réservée aux isolats portant simultanément ces deux SNPs et non seulement la polymorphisme mutT2 G1286766C.

Les autres *modern Beijing* (n = 65), sont divisés en *asian african* 1 (n = 2), *asian african* 3 (n = 5), *asian african* 2 (AAfr2, n = 16), *asian african* 2-RD142 (AAfr2-RD142, n = 1), *central asian* (n = 13), Europe/Russia B0/W148 outbreak (n = 2) et il restait 26 isolats *modern Beijing* non classifiés, qui ne correspondaient à aucune définition de sous-lignée moderne décrite, et qui n'ont pas été étudiés plus avant dans cette étude. Aucun isolat Pacific RD150 n'était inclus dans cette étude. Nous présentons un schéma

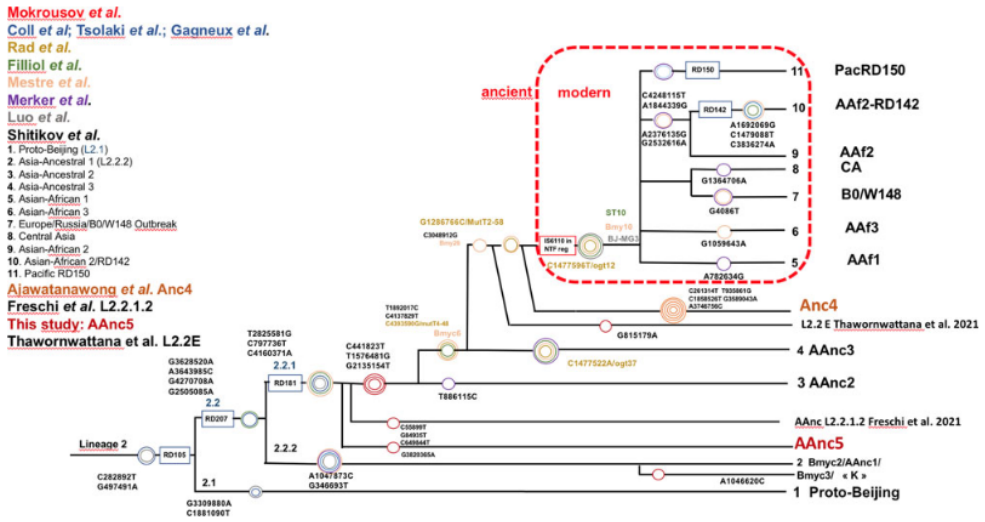


FIGURE 4.3 – Dendrogramme unifié de la lignée 2 de Mtb représentant les sous-lignées actuelles de L2 avec certains de leurs SNPs ou marqueurs génétiques. Le code couleur tente de se superposer à chaque auteur, la taille des cercles est arbitraire ; cet arbre tente de fournir un schéma évolutif unifié simplifié mais ne prétend pas représenter la diversité complète de L2 (redessiné et amélioré à partir de Shitikov et al., 2017)..

de classification L2 unifié et amélioré qui inclut la découverte récente de AAnc4, notre propre AAnc5, ainsi que le «L2.2.1.2» [88], la souche «K» (un membre de AAnc1) [113], et le L2.2.E [279], comme montré sur la Figure 3.

La réanalyse de l'ensemble des SNP phylogénétiques décrits par Shitikov et al. a confirmé tous les SNPs phylogénétiques spécifiques des sous-lignées ancestrales L2 jusqu'à la définition du L2 moderne. Le SNP mutT2 (G1286766C) est un bon marqueur phylogénétique car il est présent uniquement dans les sous-lignées modernes L2 tandis que T1892017C, C4137829T et C4393590G (mutT4) sont présents dans certaines sous-lignées ancestrales. Il est clair, d'après les résultats des SNP et de l'arbre phylogénétique présenté dans la Figure 2, que la branche AAnc4 pourrait être qualifiée d'intermédiaire, comme suggéré par l'équipe qui l'a découverte, et n'est ni une sous-lignée moderne L2 authentique ni véritablement ancestrale [4]. Inversement, les 20 isolats que nous avons étudiés du Japon, désignés par la suite comme AAnc5, remplissent tous les critères des L2 ancestraux puisqu'ils ne possèdent pas le SNP mutT2 (58) attendu ni aucune autre caractéristique de *modern Beijing*. Ils se ramifient également avant AAnc4. Dans tous les cas, la séparation nette des isolats L2 qui possèdent aucun, un ou les deux SNPs mutT2 (58) et ogt (12) confirme que ces deux marqueurs sont d'excellents marqueurs phylogénétiques. Sur la base des résultats obtenus, nous avons donc adapté le schéma de classification unifié de Shitikov et al. pour inclure certaines des nouvelles sous-lignées récemment décrites (figure 4.3). Parmi celles-ci, figurent la souche «K» (partie de AAnc1), le «L2.2.1.2», le L2.2.E [88, 277, 113], et le AAnc5 récemment désigné comme L2.2.A [277].

Description d'une nouvelle sous-lignée japonaise de l'Asia-ancestral 5 (AAnc5)

Résultats basés sur les SNPs Nous avons davantage caractérisé la sous-lignée ancestrale 5 d'Asie Japon (AAnc5) (n = 20) qui se situe entre AAnc1 et AAnc3 (figure 4.2 b et c). Ces isolats peuvent tous être définis par de nombreux SNPs partagés exclusivement. La classification basée sur les SNPs produite par le pipeline TB-Annotator sur les 190 génomes L2 est disponible dans le Tableau Supplémentaire S3. La liste de 42 SNPs exclusifs, trouvés dans 16 de ces 20 isolats, est présentée dans le Tableau Supplémentaire S4. Une observation intéressante est que toutes les souches AAnc5 possèdent une mutation non-synonyme à la position 765140 (G->C) dans rpoC. Trois autres génomes (2 dans

d'autres sous-lignées L2 et un dans une sous-lignée L4) portent également cette mutation rpoC suggérant une acquisition indépendante.

Une matrice de distance par paires entre les isolats AAnc5 a également été calculée (Tableau Supplémentaire S5) ; le pipeline actuel calcule systématiquement la distance SNP intra-branche pour les clusters ou sélections d'intérêt de moins de 100 SRAs (résultats non montrés). La distance par paires entre les échantillons AAnc5 montre un minimum de 166 SNPs (entre DRR157280 et DRR157281) et un maximum de 439 SNPs (entre DRR130203 et DRR034366) parmi les isolats AAnc5, excluant ainsi une transmission récente (Tableau Supplémentaire S5). En supposant un taux de mutation SNP moyen de 0,3 par an par génome, ces souches pourraient avoir divergé il y a environ 250 à 600 ans de leur MRCA. Si nous acceptons la suggestion du moment de l'émergence ou de l'expansion d'AAnc4 autour du 7ème siècle en Thaïlande, (début de Chiangrai), alors AAnc5 pourrait avoir été introduit au Japon plus tôt, en accord avec les informations archéologiques [4, 274, 273]. Une seconde liste de 46 SNPs partagés exclusivement, partagés entre les deux isolats les plus éloignés sur une sous-branche spécifique d'AAnc5 (DRR034381 et DRR130203, distance par paires : 417 SNP) est également présentée dans le Tableau Supplémentaire S5.

Spoligotypage in silico et reconstruction de la structure du locus CRISPR de AAnc 5 en utilisant TB-annotator, le nombre de copies de IS*6110* et les sites d'insertion Les vingt isolats AAnc5 étudiés ont montré 6 motifs de spoligotype différents tels que reconstruits par CRISPR-builder, ce qui était un résultat inattendu pour une sous-lignée L2 (Tableau Supplémentaire S6) ; la plupart de ces motifs ont été précédemment décrits dans la base de données SITVIT (SIT1, SIT190, SIT269, SIT1364, SIT1674), cependant un est resté non défini comme SIT"X". Aucun SNP n'a été trouvé dans les *spacers* et les répétitions, mais trois isolats ont présenté des duplications : une duplication de sp65 pour DRR034478 et SRR130160, et de sp50 pour DRR034476 (Tableau Supplémentaire S6). La phylogénie qui peut être dérivée de la reconstruction de la structure CRISPR-Cas a confirmé les résultats des SNPs : elle révèle des suppressions sporadiques de gènes cas, Rv2807c, Rv2808c et Rv2813c chez certains isolats (Tableau Supplémentaire S7).

Les souches AAnc5 portaient de 14 à 22 copies de IS*6110*, et deux copies spécifiques ont été trouvées dans presque tous ces isolats L2 et non dans d'autres sous-lignées L2 : une copie a été trouvée à la position 1724419 dans Rv1527c (trouvée chez 15 de ces isolats) et la seconde à la position 2041756 (trouvée chez 19 de ces isolats) (Tableau Supplémentaire S6). DRR034455, DRR034471 et DRR034476 présentaient la même structure CRISPR, cependant portaient différents gènes manquants (voir paragraphe suivant). En utilisant TB-Annotator, 14 des isolats AAnc5 ont été prédits comme sensibles aux médicaments et quatre portaient des mutations de résistance mono-résistante, deux étaient des MDRs (Tableau Supplémentaire S8).

Gènes manquants Six isolats de TB parmi les 20 AAnc5, en plus de présenter des délétions classiques (RD105, RD207, RD181 et PhiRv1), contenaient des gènes spécifiquement absents : par exemple, DRR034363 avait les gènes Rv1081 à 1084c supprimés, DRR034416 était dépourvu de Rv1523 à Rv1526c. (Tableau Supplémentaire S7). Ces délétions confirment que les génomes de *Mycobacterium tuberculosis* complex phylogénétiquement liés peuvent parfois contenir des délétions spécifiques à la souche, dues à des événements de recombinaison.

Calcul du nombre de copies de VNTR in silico en utilisant CAST et comparaison avec d'autres isolats issus d'études précédentes Aucune signature VNTR spécifique de 15 + 9 n'a pu être obtenue par typage VNTR *in silico* en utilisant CAST pour aucun des isolats AAnc5 [313, 6]. ETRC, QuB26 et QuB4156 n'ont jamais pu être

prédits *in silico*. Selon la qualité de SRA, entre 6 et 20 VNTR pouvaient être prédits (Supplementary Table S9). Les résultats VNTR ont montré une légère variation entre les isolats ; onze loci VNTR étaient invariants dans cette collection (MIRU04, MIRU10, MIRU16, MIRU20, Mtub29, Mtub30, ETRB, MIRU24, MIRU27, Mtub34, MIRU39) tandis que neuf loci présentaient des variations (MIRU02, MIRU40, Mtub21, QuB11b, ETRA, MIRU23, MIRU26, MIRU31, Mtub39). En comparant avec une étude VNTR approfondie réalisée précédemment, il a été démontré que AAnc5 appartenait à M10 ou M37 respectivement trouvés en Russie et à Singapour [187]. En comparant avec un ensemble de 5 isolats japonais de référence (A05N056, ID381, 4558, 4994, 4991/M) décrits comme représentant les principales sous-lignées L2 trouvées au Japon, ID381 partageait le même nombre de copies VNTR avec AAnc5 sur 9 loci (Supplementary Table S9) [132, 292]. En comparant les résultats VNTR *in silico* avec les résultats VNTR précédents issus d'études publiées en Corée, sur la souche «K», connue pour appartenir à AAnc1, nous avons trouvé une similitude relativement faible [140, 113] (Supplementary Table S9).

Comparaison entre les résultats de TB-annotator et du serveur CAST concernant les tests prospectifs de sensibilité aux médicaments et les résultats de spoligotypage. Lors de la comparaison des résultats des tests de sensibilité aux médicaments obtenus en utilisant soit le pipeline TB-Annotator soit CAST, ils étaient identiques (Tables Supplémentaires S8 et Table S9). Des résultats identiques ont également été obtenus sur la reconstruction classique du spoligotype au format de 43 *spacers*, avec une petite divergence encore inexpliquée sur un seul spacer d'un seul isolat, DRR034366, pour lequel CAST a prédit SIT250 tandis que TB-Annotator a prédit SIT290 (Table Supplémentaire S9).

AAnc5 est identique au cluster de souches ancestrales L2 endémiques G3 au Japon. En comparant les listes de SNPs, nous avons trouvé que le cluster de souches AAnc5 de la province de Tochigi partageait 36 des 42 SNPs avec la souche de référence G3 :ID381 trouvée à Kobe et Osaka (Tableau Supplémentaire S4). Seuls 6 SNPs (C587945T, G765140C, G1202113A, AGGGAG1476812A, G3148446C et G3820365A) n'étaient pas trouvés dans la souche ID381. Nous avons conclu que les souches de Tochigi étaient très probablement historiquement liées au groupe G3 de Kobe et Osaka décrit en 2006 à travers l'isolat de référence ID381. En conséquence, nous proposons de retenir les SNPs communs décrits par Wada et al. et cette étude comme caractéristiques de AAnc5 pour correspondre à la nomenclature de Shitikov et al. Nous avons positionné à la fois AAnc5 et AAnc4 dans l'arbre schématique global L2 [245, 4]. En approfondissant la liste comparative de SNPs entre notre étude et l'ancienne étude Kobe-Osaka, nous avons trouvé que DRR034489 était l'isolat le plus proche de la souche de référence G3 ID381 partageant 40 SNPs supplémentaires exclusifs au groupe G3, tandis qu'un autre cluster de 3 génomes était plus distant mais partageait 15 SNPs supplémentaires avec ID381 (résultats non montrés). Comme mentionné ci-dessus, deux génomes très éloignés, DRR034381 et DRR130203 (distance de 417 SNPs en paire) partageaient également 46 SNPs supplémentaires qui n'étaient trouvés nulle part ailleurs (Tableau Supplémentaire S5).

4.1.4 Discussion

Nous décrivons dans ce travail une sous-lignée ancestrale endémique historique de L2 basée sur des échantillons collectés dans le centre du Japon, préfecture de Tochigi, ancienne province de Shimotsuke, que nous avons nommée AAnc5. Cette sous-lignée est fortement liée au groupe G3 japonais défini en 2012 [292] et supposée être nommée L2.2.A dans une revue récente [277]. Nos résultats renforcent la pertinence phylogénétique de cette sous-lignée dans l'histoire évolutive globale de L2 et montrent qu'elle a été transmise historiquement dans plusieurs villes japonaises. La chronologie de l'émergence de cette

sous-lignée par rapport aux autres sous-lignées L2 a été positionnée dans le schéma de diversification de L2 de Shitikov. Sa position par rapport aux lignées précédemment décrites a clairement montré qu'elle devrait être qualifiée d'ancestrale selon la définition actuelle de cette terminologie et divergée des autres sous-lignées de Beijing peu après AAnc1.

La tuberculose est très ancienne dans l'histoire de l'humanité, cependant il est encore impossible de dater définitivement son émergence en Asie [166]. L'histoire précoce des épidémies de TB au Japon pourrait être liée aux migrations de personnes du 5ème siècle avant JC au 3ème siècle après JC [273, 274]. La tuberculose était connue pour être présente dans l'ancien temps au Japon sous le nom de rôga, qui était utilisé en médecine chinoise [279]. Un lien entre cette maladie et la TB, telle que connue en médecine occidentale, a été décrit en 1857 par un médecin, Ôgata Kôan, 3 ans avant l'ouverture du Japon [279]. Il existe de nombreuses traces dans les textes d'histoire médicale chinoise d'une maladie qui peut être identifiée comme la tuberculose [14]. L'historien Fan Xingzhun a donné quelques indices dans son étude succincte sur l'histoire de cette maladie en Chine [82]. Son analyse des sources chinoises l'amène à souligner que de nombreux termes associés à des symptômes évocateurs de la tuberculose apparaissent dans des sources anciennes. Le Classique des montagnes et des mers (Shanhai Jing 山海經, 4ème–3ème siècle avant JC) décrit un remède pour guérir le «luo 癧», c'est-à-dire les scrofules. Cependant, Fan Xingzhun admet que ces symptômes peuvent ne pas être spécifiques. Il note que d'autres termes évocateurs de la maladie apparaissent dans d'autres sources anciennes telles que le Classique de la poésie (Shijing 詩經, recueil de textes s'étendant du 11ème au 5ème siècle avant JC) ou dans le Mengzi (孟子, 4ème siècle avant JC), que Su You 蘇游, un auteur de la dynastie Tang (618–907), considère comme des synonymes. Fan Xingzhun souligne que ces termes (zhai 瘵, mo 瘼, chuanshi 傳尸 (littéralement : cadavre qui transmet), shoubing 瘦病 (littéralement : maladie de la minceur), zhuanzhu 轉注, fulian 伏練, guzheng 骨蒸 (littéralement : os remplis de vapeur chaude ou os chauds) sont souvent associés à une description d'états de fatigue extrême et d'émaciation dans les premiers dictionnaires et livres médicaux. L'idéogramme «zhai 瘵», en particulier, qui est devenu l'un des plus populaires pour décrire, entre autres, la tuberculose, est défini dans les premiers dictionnaires comme une maladie dont la caractéristique principale est la faiblesse, et, plus tard, comme une maladie qui conduit le patient à «se tourner et se retourner sans repos» jusqu'à «mourir de faiblesse». Bien que Fan Xingzhun convienne que tous ces termes ne correspondent pas nécessairement à la tuberculose, l'historien trouve néanmoins très probable que les preuves qu'il a trouvées dans un livre du 14ème siècle décrivent un cas de tuberculose pulmonaire : «Sous la dynastie Song (d.960-1127), Shi Dezun, lorsqu'il avait 50 ans, a contracté la maladie de l'épuisement et de la perte de poids, il se tournait et se retournait sans repos jusqu'à devenir très émacié» [82]. De plus, Fan Xingzhun émet l'hypothèse que la prévalence de la tuberculose était faible dans l'ancien temps. Cependant, dans l'Ancien Livre des Tang (Jiu Tangshu, 舊唐書 941 après JC), il est rapporté que : «sous le règne de Wude (618–626), à Guanzhong (province du Shaanxi), beaucoup ont la ‹maladie des os chauds› 骨蒸», un témoignage qui plaide en faveur d'une épidémie de tuberculose. Comme le terme «maladie des os chauds» (骨蒸) est déjà mentionné dans le Canon médical «Le Canon de médecine de l'empereur Jaune» (Huangdi Neijing 帝內經, 2BC-2AC), Fan Xingzhun s'interroge sur les possibilités que la tuberculose ait déjà pu être épidémique à cette époque [82].

L'exploration approfondie de l'histoire phylodynamique complexe de toutes les lignées de MTBC a été rendue possible grâce à l'utilisation du pipeline TB-Annotator qui analyse plus de 50 000 caractères (au moment de cette étude) incluant des séquences répétées et des SNPs [199, 88, 285, 139, 270, 178]. Avec un nombre croissant de SRAs disponibles publiquement, il devient possible de démêler tous les fils entre un événement historique ancien et récent qui a façonné la pandémie actuelle de TB, et de comprendre sa relation avec les migrations de populations anciennes/modernes [166, 212, 194].

Parmi l'ensemble des caractéristiques de AAnc5, nous décrivons une mutation non synonyme dans rpoC. Il est bien connu que les mutations rpoC sont des mutations compensa-

toires principalement trouvées dans des isolats épidémiologiquement réussis qui contiennent également des mutations spécifiques de rpoB [47, 111, 71]. La mutation rpoC pourrait être un trait adaptatif qui explique le succès épidémiologique des isolats MDR-TB L2 dans des épidémies telles que l'épidémie d'Asie centrale ou l'épidémie B0/W148 en Russie [181, 163] et le succès épidémiologique chez les prisonniers géorgiens [111]. Dans le clade L2 d'Asie centrale, les mutations rpoC peuvent survenir dans l'ensemble du gène et ces SNPs sont trouvés en épistasie avec des mutations rpoB [181, 47, 71]. Dans notre étude, la mutation rpoC, à l'exception des six isolats les plus récents, a été trouvée dans des isolats sensibles aux médicaments ; il est donc difficile de considérer une telle mutation comme adaptative ou compensatoire, mais plutôt comme un marqueur phylogénétique de AAnc5 [111]. Étant donné que ce SNP n'a pas été décrit dans le groupe G3 de Kobe et Osaka [292], il pourrait être intéressant à l'avenir d'essayer de rechercher s'il existe une différence statistique significative entre l'émergence de la résistance aux médicaments dans un groupe ou l'autre groupe de clusters d'isolats. Un article récent sur les prisonniers en Géorgie a montré que les mutations compensatoires et l'incarcération des patients étaient deux facteurs indépendants associés à l'augmentation de la transmission, créant une «tempête parfaite» pour la transmission de la MDR-TB [111]. Les conséquences physiologiques précises de cette mutation rpoC non synonyme dans des isolats sensibles aux médicaments n'ont pas été étudiées dans cette étude mais pourraient également avoir des conséquences fonctionnelles [47]. L'évolution de L2 comprend l'histoire évolutive précoce de AAnc5 et pourrait être liée à des effets mutateurs encore inconnus spécifiques de certaines sous-lignées L2 [218].

La datation de la diversification des souches peut également donner des indices sur les conditions sociales et démographiques qui ont favorisé les épidémies passées. L'étalonnage de l'horloge moléculaire est difficile selon les échantillons, les cadres temporels et les lignées (entre 0,04 et 2,2 SNPs par génome par an) [177]. La datation peut être confortée si une corrélation adéquate existe entre les faits historiques, génomiques, épidémiologiques et démographiques [194]. Dans une étude similaire antérieure, nous avons testé trois scénarios pour dater le MRCA d'une sous-lignée L4.2 survenant au Japon et en Turquie, et confronté des résultats historiques, anthropologiques, de génétique humaine, paléopathologiques et génomiques [224]. Ici, selon l'horloge moléculaire consensuelle de la TB à moyen terme, le cadre temporel de coalescence des repères de la sous-lignée AAnc5 pourrait être entre 280 et 310 ans avant le présent pour les isolats les plus proches, et 760–800 ans avant pour les plus éloignés. Nous avons basé cette estimation sur les taux de mutation de L4 et non de L2 et sommes donc très prudents.

La province de Tochigi est célèbre pour sa grande mine de cuivre d'Ashio, dont l'exploitation a commencé au début du 17ème siècle. L'une de nos estimations de la date d'expansion de AAnc 5 est compatible avec l'ouverture de la mine à Ashio. La mine d'Ashio pourrait avoir été un lieu d'expansion et de diversification de AAnc5. Examiner la dynamique de propagation historique du groupe AAnc5 au Japon pourrait être fait à l'avenir avec l'aide d'enquêteurs japonais en examinant des gradients plus fins de toutes les sous-lignées ancestrales L2 prévalentes trouvées au Japon.

Une autre origine plus récente des AAnc5 japonais pourrait être l'importation de clones ancestraux L2 par des travailleurs forcés de Corée ou de Chine qui ont été contraints de travailler dans l'industrie minière de 1939 à 1945 [279]. En effet, 300 000 Coréens et 38 935 travailleurs chinois, principalement des hommes, ont été forcés de travailler pour le Japon pendant ces années [279]. Cependant, étant donné le nombre élevé de SNPs accumulés à l'intérieur de AAnc5, cette hypothèse semble peu probable. Une limitation de cette étude est que nous n'avons pas pu enquêter sur la distribution potentielle des isolats G3/AAnc5 dans tout le Japon, cependant notre preuve formelle que G3 et AAnc5 sont liés et profondément enracinés est la première perspective sur une histoire complexe. Les études futures pourraient essayer de disséquer davantage, en utilisant une combinaison de statistiques récentes sur le total des cas épidémiques, en combinaison avec la caractérisation génomique et la distribution géographique, l'histoire globale et locale des lignées ancestrales L2 au

Japon [206].

L'existence d'une diversité relativement élevée d'isolats spécifiques à Tochigi et à Kobe et Osaka rappelle une transmission historique restreinte à une zone circonscrite. Bien sûr, une telle image d'une épidémie endémique passée est plus facilement observée dans des cadres insulaires comme cela a été récemment montré en Nouvelle-Zélande chez les Maoris qui hébergent une sous-lignée spécifique «CS» (Colonial S-type) L4.4 [194]. Les îles sont d'excellents cadres pour distinguer les espèces endémiques des espèces importées et l'histoire de la tuberculose est définitivement liée à la migration humaine et à l'histoire démographique locale et mondiale [252]. Selon la revue la plus récente sur la structure de la population L2 basée sur plus de 5000 génomes, AAnc5 ou L2.2.A est le clade le plus basal de L2.2.1 et comprend des isolats presque entièrement du Japon. La structure profondément ramifiée de L2.2.A suggère une souche endémique auparavant non reconnue [277]. Étant donné que le Japon est une île, et un pays avec un taux de prévalence de la tuberculose très faible actuellement (13 pour 100 000 en 2017), c'est un excellent cadre pour identifier les événements historiques liés aux épidémies de tuberculose passées [224]. Le nombre croissant de génomes disponibles permet de découvrir de plus en plus de sous-lignées L2, cependant leurs relations historiques et épidémiologiques intimes restent à étudier plus en détail [88, 277].

4.1.5 Conclusion

Grâce à TB-Annotator, nous avons cartographié une sous-lignée ancestrale endémique L2 du Japon sur la phylogénie mondiale du *Mycobacterium tuberculosis* complex, et nous l'avons désignée sous le nom de *asia ancestral 5* (AAnc5). Nous avons également démontré qu'elle était liée au groupe G3 précédemment décrit à Kobe et Osaka, désormais désigné sous le nom de L2.2.A. Cette sous-lignée apparaît désormais aux côtés de certaines parmi les plus récentes dans un schéma évolutif unifié. AAnc5 possède de nombreux caractères spécifiques qui lui permettent de se distinguer de toutes les autres sous-lignées ancestrales décrites jusqu'à présent dans L2. Cette découverte ouvre de nouvelles voies de recherche pour explorer l'histoire de L2 en Asie du Sud-Est.

4.2 Vers une meilleure compréhension de l'histoire évolutive de *Mycobacterium tuberculosis*

4.2.1 Introduction

Bien que de nombreux progrès aient été réalisés dans la connaissance de la phylogénie du MTBC grâce à de nombreuses études comparant plusieurs centaines ou milliers de génomes séquencés, il reste encore de nombreuses zones d'ombre que l'augmentation de la quantité de données analysée devrait nous aider à éliminer. Parmi ces zones d'ombre de la phylogénie, nous avons choisi d'illustrer dans ce chapitre les cas suivants. Tout d'abord, la lignée 6 [44] n'a été explorée que récemment par Coscolla et al. [62], mais avec un nombre restreint d'échantillons (n = 338), n'incluant pas toutes les données actuellement disponibles et n'incluant pas de groupes externes, de sorte que plusieurs SNPs de classification manquent de spécificité. La lignée 4 [44, 266], en revanche, a été fortement séquencée mais est actuellement hétérogène, avec certaines sous-lignées fortement représentées et subdivisées (par exemple 4.1) et d'autres mal caractérisées (4.5 et 4.7). Enfin, aucune sous-lignée d'Éthiopie L7 [44] n'est actuellement définie.

L'objectif de ce chapitre est d'illustrer ces imperfections et de montrer qu'avec un nouvel outil ad hoc, à savoir le TB-annotator précédemment présenté, nous avons les moyens d'éclaircir ces zones d'ombre, en évaluant la qualité des SNPs actuellement sélectionnés pour la définition des lignées, et en recherchant de nouveaux clades partageant des caractères exclusifs.

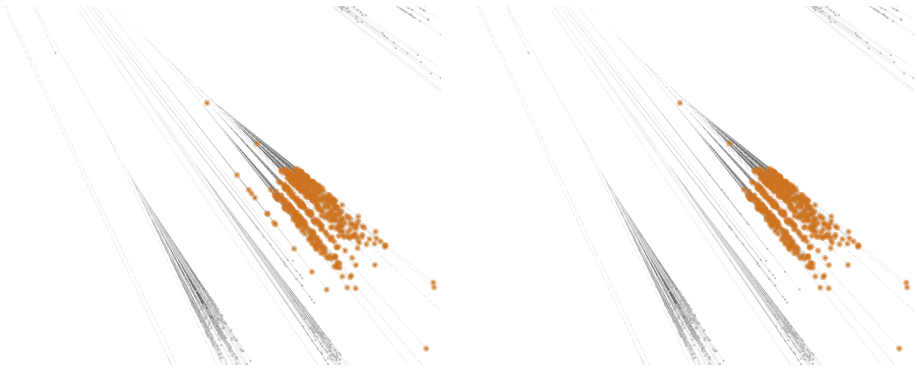
4.2.2 Résultats de TB-annotator, à partir de la version à 15 901 souches

L'analyse de 15 901 souches par notre pipeline a permis de détecter 180 RDs et plus de 300 000 SNPs de bonne qualité. Cette liste contenait les SNPs proposés par Coscolla [44] pour définir les sous-lignées L6, ceux de Stucki [266] pour les sous-lignées L4, et les SNPs de Coll [44]. Cela nous a permis de remettre en question la qualité de ces SNPs en fonction des 15 901 souches. Nous avons également profité de l'occasion pour examiner deux propositions de caractérisation de lignées plus récentes, celles de Napier [200] et de Freschi [87].

Concernant L6, rappelons d'abord qu'elle a été définie par le SNP C1816587G dans Coll et al. (2014), mais que l'équipe n'a proposé aucune sous-lignée. Coscolla et ses collaborateurs, pour leur part, ont identifié trois sous-lignées, numérotées 6.1, 6.2, 6.3, chacune ayant à son tour trois sous-lignées. Bien qu'ils aient proposé des SNPs pour les sous-lignées 6.1.1 à 6.3.3, ils n'ont proposé aucun SNP pour L6, ni pour 6.1 à 6.3. Napier et al., quant à eux, ont proposé 10 nouveaux SNPs pour définir L6, mais n'ont pas retenu celui de Coll. Ils n'ont également défini aucune sous-lignée. Enfin, Freschi est revenu à la définition de L6 par Coll (C1816587G), sans sous-lignée.

Le SNP de Coll pour L6 a été identifié dans 609 souches dans l'outil TB-annotator, et cette sélection a montré deux variants exclusifs. Cependant, en regardant dans le détail, les souches sélectionnées par ce SNP comprennent à la fois L6 et L9, voir figure 4.4a. De notre côté, nous avons trouvé que le SNP G41241A permet de définir exclusivement L6 (en excluant L9) : il est présent dans 599 souches qui partagent 78 variants exclusifs (figure 4.4b). En explorant plus en profondeur la proposition de Coscolla, il semble que 6.3.1 soit bien défini, mais pas 6.3.2 et 6.3.3 : le nombre de souches présentant les différents SNPs, présentés comme spécifiques, varie du simple au double selon le SNP choisi (plusieurs SNPs sont proposés pour la même sous-lignée). Concernant le 6.2 de Coscolla, cette sous-lignée était bien soutenue par TB-annotator mais le SNP de 6.2.2 apparaissait également sur 6.2.1 et 6.2.3. Des observations similaires ont été faites pour 6.1 : les SNPs de 6.1.1 et 6.1.3 sont bien soutenus par TB-annotator, mais 6.1.2 était mal défini, incluant 2 souches de 6.1.3. Enfin, dix isolats de 6.1 n'appartiennent à aucune sous-lignée, bien qu'ils semblent former un clade.

Concernant L7, les définitions de Freschi et Napier correspondent à la définition de Coll



(a) Le SNP C1816587G de Coll (b) Notre SNP G41241A

FIGURE 4.4 – Le SNP de L6 de Coll englobe à la fois L6 et L9. Comme le montre la figure de gauche, le SNP proposé par Coll pour définir L6 inclut également des souches en dessous du clade regroupant tous les L6. Après analyse plus approfondie, il s'avère que ces souches inférieures appartiennent à L9. Notre SNP sépare L6 de L9, comme le montre la figure de droite.

du SNP G1137518A. Cependant, ce SNP n'atteint que 97% d'exclusivité. En revanche, TB-Annotator a révélé 836 variants qui sont à 100% exclusifs. C'est le cas de G1960C, que nous proposons comme le meilleur marqueur de L7. De plus, aucune sous-lignée de L7 n'est actuellement définie, alors que notre interface a mis en évidence qu'au moins 3 sous-lignées pourraient être définies. La première comprend au moins 10 souches et présente diverses variants exclusifs, dont T109162C. De même, G160716A définit un groupe de 18 souches et G1932863T un groupe d'au moins deux souches. Des études plus approfondies sur la qualité de ces SNPs et des clades associés sont nécessaires pour confirmer leur pertinence.

Concernant la lignée 4, comme nous l'avons vu, certaines sous-lignées sont bien plus caractérisées que d'autres. Par exemple, 4.7 n'a aucune sous-lignée dans les phylogénies actuellement proposées, et Stucki, Freschi, et Napier utilisent tous la définition de Coll (C4249732A). En utilisant TB-annotator, nous avons d'abord trouvé une meilleure caractérisation de 4.7 (C10741G), et de nombreuses sous-lignées semblent pouvoir être définies à partir de SNPs. Un schéma potentiel est reproduit dans la figure 4.5, qui nécessitera une confirmation dans des travaux futurs. Il en va de même pour 4.5, dont la liste des SNPs de sous-lignées pourrait être celle du tableau 4.1.

FIGURE 4.5 – Une proposition de définition des sous-lignées pour L4.7. Cette subdivision originale de L4.7 en sous-lignées a été obtenue dans TB-annotator en recherchant des clades significatifs ayant au moins un caractère exclusif.

4.2.3 Discussion

La caractérisation la plus détaillée de la lignée 6 est pour le moment celle proposée par Coscolla. Les 3 principales sous-lignées qu'ils définissent semblent être correctes selon notre

Lignée	SNP
451	275523
452	275624
453	212629
454	570543
4511	85961
4512	501530
4513	3879070
45131	39725
45132	255444
4531	316970
4532	128589
45321	4294863
45322	1853597
4541	194738
4542	4314
45421	384635
45422	385778
4543	2331267

TABLE 4.1 – SNPs possibles pour une phylogénie de 4.5. Ces SNPs ont été trouvés en utilisant TB-annotator : pour chaque clade significatif dans l'arbre, nous avons vérifié s'il existait au moins un caractère exclusif. Si c'est le cas, nous avons proposé une définition de sous-lignée, avec le SNP exclusif comme caractéristique.

étude, ainsi que l'existence de sous-lignées pour 6.1 à 6.3. Cependant, il est regrettable qu'il n'y ait de caractéristiques que pour le troisième niveau de classification, et que ni L6 ni 6.1 à 6.3 n'aient de SNP ou liste de SNP de définition.

Avec notre interface, de nombreux candidats (variants exclusifs) apparaissent, comme C3887907A pour le 6.1, G4393A pour le 6.2, et C98536T pour le 6.3. Ces candidats devront être étudiés plus avant pour une meilleure compréhension de L6. De même, il semble que de meilleurs SNPs puissent être définis pour chacune des sous-lignées 6.1.1 à 6.3.3, réduisant ou éliminant les faux positifs (souches en dehors du clade qui ont tout de même le SNP) et les faux négatifs (souches dans le clade sans le SNP). Enfin, il nous semble que ces sous-lignées peuvent à leur tour être divisées en sous-lignées, ce qui fera l'objet d'études futures.

L7, pour sa part, est clairement sous-étudiée. Nous avons vu que le SNP qui la définit depuis presque dix ans n'est pas optimal. De plus, il semble tout à fait possible de définir des sous-lignées dans L7, mais cela nécessiterait probablement un plus grand nombre de génomes pour être confirmé.

Enfin, nous avons montré que les lignées 4.5 et 4.7 sont actuellement mal caractérisées (présence de faux positifs et faux négatifs) et mal décomposées (pas de définitions de sous-lignée), bien que de telles définitions soient tout à fait possibles.

L'objectif du travail de ce chapitre n'était pas de proposer une nouvelle phylogénie, mais de soulever de nombreux points montrant, d'une part, que les caractérisations actuelles sont perfectibles, et d'autre part, que nous pouvons probablement aller plus loin dans la définition de certaines sous-lignées avec de nouveaux outils, tels TB-annotator, exploitant l'ensemble des données de séquençage actuellement disponibles. Ces améliorations, réalisées par exemple dans [230, 108], sont rendues possibles par le très grand nombre de génomes dont nous disposons désormais, mais nécessitent l'utilisation de tels outils ad hoc pour gérer de telles quantités de données et produire les analyses nécessaires à ces nouvelles définitions.

4.3 Identification d'une nouvelle lignée : *Mycobacterium africanum* lignée 10

4.3.1 Introduction

TB-Annotator a permis d'affiner la connaissance phylogénétique des lignées. Mais il se pose aussi la question de savoir si toutes les lignées du MTBC ont été découvertes, ou s'il reste encore des lignées non décrites dans la littérature. Si tel est le cas, ces lignées sont nécessairement petites et exotiques, compte tenu de la taille des ensembles de données des dernières études de référence [87, 200]. À cette fin, nous avons interrogé la plateforme TB-Annotator, en recherchant tous les génomes qui ne contenaient pas le SNP de référence pour la lignée 1, ou la lignée 2, etc., et ce pour toutes les lignées connues, humaines ou animales. Par «SNP de référence», nous entendons les SNPs donnés comme référence dans la littérature, que nous avons également validés à l'aide de TB-annotator.

Aux génomes résultants de cette requête, nous avons ajouté des génomes représentants la diversité de chaque lignée connue, pour servir d'arbre de référence et de base de connaissances pour l'arbre phylogénétique que nous avons ensuite construit et chargé dans TB-annotator. De longues branches sont apparues, correspondantes à des génomes en dehors du complexe MTBC et incorrectement étiquetés *M.tuberculosis* sur le NCBI. Après avoir nettoyé la sélection de ces souches, nous avons construit un deuxième arbre, que nous avons à son tour chargé dans TB-annotator. Un certain nombre de souches étaient proches de la racine, qui, après investigation, se sont avérées être des génomes problématiques (mauvaise qualité, pollution, etc.). Un troisième arbre, reconstruit sans ces souches et ensuite visualisé, a révélé toutes les lignées connues et deux souches clairement séparées du reste, que nous avons ensuite identifiées comme la nouvelle L10.

Une telle découverte n'a été rendue possible que par une analyse massive de toute les données de séquençage disponibles publiquement, et par la combinaison de requêtes complexes sur les marqueurs génétiques extraits ainsi que la visualisation de leur lien avec l'arbre phylogénétique.

Ce chapitre décrit cette lignée nouvellement identifiée et jusqu'à présent rare du complexe *Mycobacterium tuberculosis*, la L10 (proposée), présente en Afrique centrale. Comme nous le verrons ci-après, cette lignée est caractérisée par une nouvelle RD, des insertions IS*6110* et 243 SNPs, y compris gyrA G7901T, recN C1920096T et dnaG C2621730T. L10 représente un clade frère de L6, principalement trouvé en Afrique de l'Ouest, et de L9, spécifiquement en Afrique de l'Est, et révèle une pièce potentiellement manquante dans l'histoire évolutive et les migrations de *M. africanum*. Ces découvertes étendent la diversité connue de *M. africanum* en Afrique.

4.3.2 Objectif de l'étude

La vision traditionnelle d'une diversité restreinte parmi les agents bactériens responsables de la tuberculose humaine et animale est en cours de révision grâce à l'utilisation généralisée du WGS. En plus de *Mycobacterium canettii*, qui représente des lignées exceptionnelles, non clonales et anciennes de bacilles tuberculeux en Afrique de l'Est, plusieurs lignées jusqu'alors inconnues du complexe *M. tuberculosis* ont été identifiées en Afrique au cours de la dernière décennie. La lignée 7 (L7) du complexe *M. tuberculosis* a été découverte dans la Corne de l'Afrique et la L8 dans la région des Grands Lacs africains [93, 203]. La lignée *M. africanum* L9 n'a été trouvée qu'à Djibouti et en Somalie. En revanche, deux autres grandes lignées affiliées à *M. africanum*, contribuant de manière significative à la charge de la tuberculose, L5 et L6, se retrouvent principalement en Afrique de l'Ouest [60]. La voie évolutive reliant l'Afrique de l'Est et de l'Ouest dans l'histoire du bacille reste floue. Nous décrivons ici une lignée sœur nouvellement identifiée de L6 et L9, associée à l'Afrique centrale, et discutons des implications pour déterminer l'histoire évolutive des lignées apparentées *M. africanum* L5, L6 et L9.

4.3.3 La découverte d'une nouvelle lignée

Nous avons utilisé la plateforme TB-Annotator, développée dans le cadre de cette thèse pour intégrer les données WGS, provenant cette fois-ci de 102 001 isolats du complexe *M. tuberculosis* accessibles dans le domaine public avec le NCBI.

Les SNP provenant d'un ensemble exploratoire de 15 699 isolats, principalement d'origine africaine, ont été utilisés pour construire un arbre phylogénétique. Notre analyse a alors permis d'identifier une lignée sœur de *M. africanum* L6 et L9, se ramifiant entre ces lignées et la lignée animale A1 (La_A1) [60]. La lignée nouvellement identifiée est représentée par seulement deux génomes : ERR2707158, obtenu à partir d'une souche isolée en 2008 d'un patient résidant à Kinshasa, en République Démocratique du Congo (RDC), désormais incorporée sous la référence ITM-501386 (CT2008–03226) dans les collections coordonnées de micro-organismes de l'Institut de Médecine Tropicale (Anvers, Belgique) ; et ERR2516384, obtenu à partir d'une souche isolée en Belgique en 2013 (V. Mathys, communication personnelle, e-mail, 5 juillet 2023). Les génomes de cette nouvelle lignée ne portaient aucun des marqueurs SNPs décrits dans le dernier schéma de classification des lignées du complexe M. tuberculosis [200] et aucun SNP conférant une résistance connue aux médicaments.

Pour confirmer la position phylogénétique de ces deux génomes, nous avons identifié les SNPs de 132 isolats couvrant la diversité génétique et géographique des lignées L5 et L6, et incluant des représentants de toutes les autres lignées à l'aide du pipeline Genotube (A. Le Meur, communication personnelle, e-mail, 15 septembre 2023) et TB-Profiler [213]. La reconstruction phylogénétique résultante a confirmé le regroupement de ERR2707158 et ERR2516384 dans une branche située entre L6, L9 et la lignée animale La_A1 (figure 4.6).

Les échantillons nouvellement désignés L10 partagent 375 SNPs spécifiques avec des isolats de notre ensemble de 132 échantillons sélectionnés ; parmi ceux-ci, 243 SNPs spécifiques n'ont été détectés dans aucun des 102 001 génomes inclus dans TB-Annotator. Parmi ces SNPs spécifiques, 91 sont synonymes. La distance entre les deux échantillons d'intérêt est de 382 SNPs (SNPs en dehors des régions répétitives, vérifiés manuellement en cas de discordance entre deux pipelines), beaucoup plus courte que la distance aux autres échantillons de notre sélection (minimum 1 137 SNPs ; moyenne 1 591 $\pm$ 222 SNPs).

Nous avons ensuite exploré d'autres caractéristiques des génomes pour corroborer les inférences phylogénétiques basées sur les SNPs. En plus de la délétion RD9 partagée avec la branche L5/L6 et les lignées associées aux animaux, les deux génomes L10 ne présentaient pas les délétions RD7, RD8 et RD10 [60]. Cependant, ils ne montraient pas non plus les délétions RD702 (L6/L9) ou RD713 (L5). En revanche, les deux génomes non classifiés comportaient une grande délétion spécifique de 9 134 nucléotides (Rv0613c-Rv0622) dans *M. tuberculosis* H_{37}Rv (NC_000962.3 :706602-715736), non observée dans aucune autre lignée. Ce segment incluait le couple de gènes toxine/antitoxine vapB29/vapC29. Deux autres délétions partagées englobent les gènes eis et dnaE2, pouvant potentiellement limiter la capacité d'acquérir une résistance aux aminosides [34] et affecter certaines propriétés mutationnelles [77] de ces souches *M. africanum*. Les deux génomes partagent également 4 copies IS*6110* à une position unique pour cette lignée. Dans le locus CRISPR des deux génomes L10, reconstruit à l'aide de CRISPRbuilder-TB [107], nous avons trouvé la même absence de *spacers* 7 et 9 (format spoligotype à 43 *spacers*), observée dans L6, L9 et La_A1 (voir Tableau 4.2), ainsi que l'absence de tous les *spacers* à partir des *spacers* 22 (ERR2516384) ou 26 (ERR2707158).

Les caractéristiques génétiques des souches identifiées, combinant une position phylogénétique distincte, une distance génétique par rapport à la branche L6/L9 et aux autres lignées connues de *M. tuberculosis*, des régions de délétion spécifiques, des insertions IS*6110* et des signatures de spoligotype particulières, nous ont amenés à proposer leur classification dans une nouvelle lignée désignée L10. Nous proposons trois SNP synonymes (gyrA G7901T, recN C1920096T et dnaG C2621730T) en comparaison avec la séquence de référence H37Rv NC_000962.3 dans les gènes de ménage pour identifier cette nouvelle lignée.

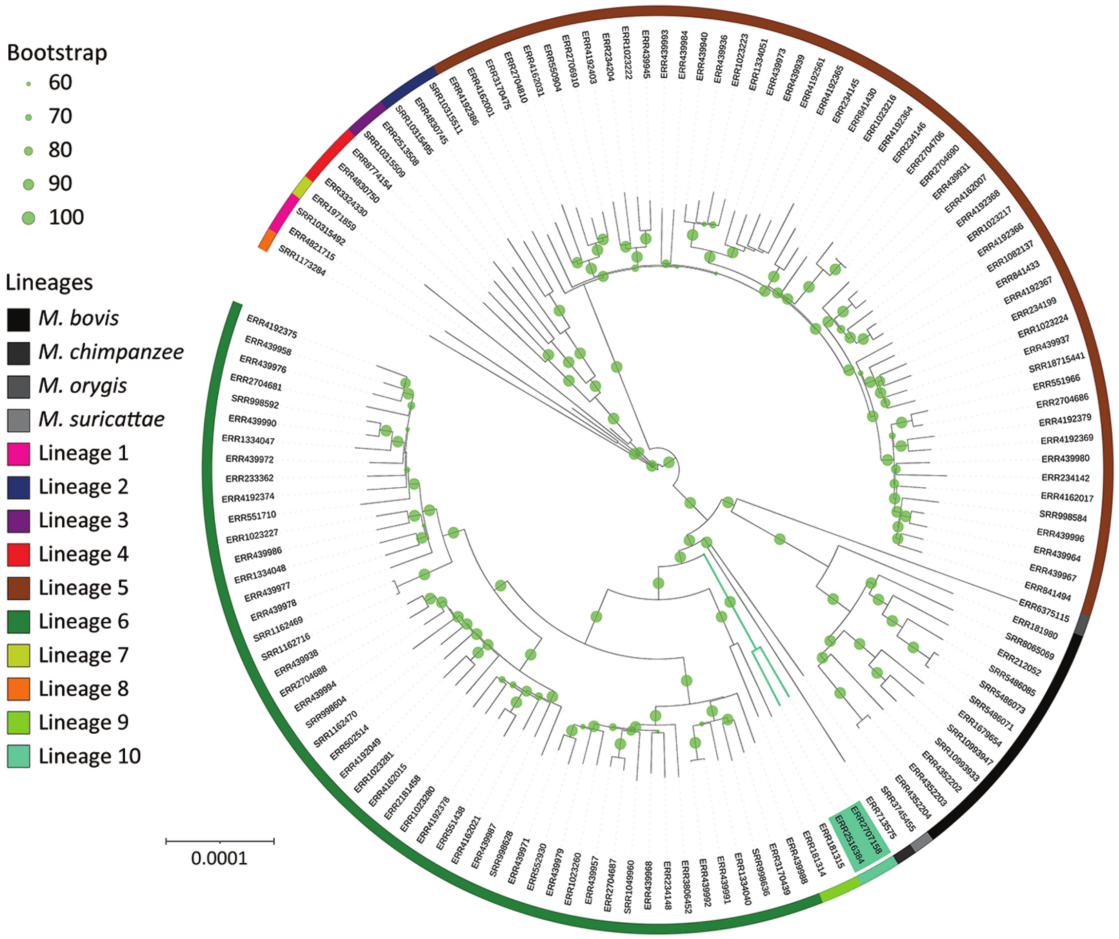


FIGURE 4.6 – Phylogénie globale de *Mycobacterium*, incluant les souches nouvellement identifiées de *M. africanum* L10 (proposition) (ombrage vert). Nous avons sélectionné des échantillons de *M. africanum* présentant la délétion RD9 et dont le pays d'origine est documenté, puis affiné notre sélection pour ne retenir qu'un représentant unique de chaque sous-lignée par pays. Cet échantillon représente la diversité génétique et géographique de *M. africanum* en Afrique. Pour cette reconstruction phylogénétique, les SNPs ont été identifiés en comparaison avec un ancêtre de *M. tuberculosis* [46] et réincorporés dans le génome complet, afin d'éviter les biais dans le modèle moléculaire ou le recours à la correction de Lewis. La phylogénie a été enracinée avec *M. canettii*, qui a ensuite été retiré pour améliorer la visualisation. Le support bootstrap a été calculé avec 100 réplicats et est indiqué lorsqu'il est >0,6. Les cercles confirment le large support de presque toutes les branches, en particulier celles de L10 et de ses branches sœurs. Le point de ramification de L10 se situe entre L9 et la lignée La_A1 regroupant les bacilles du chimpanzé et du dassie. La barre d'échelle indique les substitutions nucléotidiques par site.

Pour évaluer la circulation potentielle des souches L10 aux niveaux régional et mondial, nous avons recherché des motifs similaires de spoligotype dans SITVIT2, qui réunit des spoligotypes de plus de 110 000 isolats de 131 pays [65]. Nous avons identifié une seule instance, BEL04200301729, présentant le même motif de spoligotype que ERR2516384, représentant peut-être une troisième occurrence de L10. À noter que cette souche a été isolée en République du Congo, un pays voisin de la RDC où ERR2707158 a été collecté. Nous avons également examiné les résultats de spoligotypage issus de séquençages de nouvelle génération, collectés auprès d'environ 1 500 isolats lors d'une enquête nationale en RDC en 2016–2017, ciblés avec Deeplex Myc-TB (https ://www.deeplex.com [147]) sans détecter de motif similaire. Ainsi, les bases de données mondiales (TB-Annotator et SITVIT2) et locales suggèrent que les souches L10 sont rares au niveau mondial et

ID	Source†	No.	Spoligotype binary	SIT	isolation
L10-BEL04200301729	SITVITWEB	1	■■■■■■□■□■■■■■■■■■■■■□□□□□□□□□□□□□□□□□□□□□□□□	Orph	Republic of Congo
L10-ERR2516384	TB-Annotator	1	■■■■■■□■□■■■■■■■■■■■■□□□□□□□□□□□□□□□□□□□□□□□□	Orph	Belgium‡
L10-ERR2707158	TB-Annotator	1	■■■■■■□■□■■■■■■■■■■■■■■■■□□□□□□□□□□□□□□□□□□□□	Orph	DRC
L6_SIT181	SITVITWEB	208	■■■■■■□□□■■■■■■■■■■■■■■■■■■■■■■■■■■■□■■■■■	181	Gambia
L9	Coscolla 2021	3	■■■■■■□■□■■■□□□□□□□□□□□□□□□□□□□□□□□■■■■□■■■■■	Orph	Somalia
L9	Coscolla 2021	1	■■■■■■□■□□□□□□□□□□□□□□□□□■■■■■■■■■■□■■■□■■■■■	U	Djibouti
L9_FXX01199901706	SITVITWEB	1	■■■■■■□■□□□□□□□□□□□□□□□□□■■■■■■■■■■□■■■□■■■■■	Orph	France
NLD009501731	SITVITWEB	2	■■■■■■□■□■■■■□□□□□□□□□□□□□□□□□□□□□□□■■■■□■■■■■	Orph	Netherlands
A1_Dassie bacillus	TB-Annotator	1	■■■■■■□■□■■■□□□■■■■■□□□□□□□□□□□□□□□□□■■■□□■■■	U	South Africa
A1_*M. mungi*	https://mbovis.org, TB-Annotator	1	■■□■■■□■□■■□□□□□□□□□□□□□□□□□□□□□□□□□□□□□■■□■■■■■	SB 1960	Botswana
A1_chimpanzee bacillus	Coscolla 2013, TB-Annotator	1	■■■□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□■■■■■■■□■□□■	U	Côte d'Ivoire

*L6, L9, and both L10 samples harbor the same losses of spacer 7 and of spacer 9. DRC, Democratic Republic of the Congo; orph, orphan (single reported occurrence); SIT, spoligotype international type; U, undesignated.
†Coscolla 2021, (*3*); Coscolla 2013, (*12*); SITVITWEB, (*9*); TB-Annotator, G. Senelle, unpub. data, https://www.biorxiv.org/content/10.1101/2023.06.12.526393v1.
‡Origin unknown.

TABLE 4.2 – Motifs de spoligotype des souches nouvellement identifiées de *Mycobacterium africanum* L10 (proposition) provenant d'Afrique centrale, comparés aux souches représentatives des lignées L6, L9 et A1.

probablement restreintes à l'Afrique centrale, en dehors de quelques diffusions migratoires. La cartographie de la diversité de *M. africanum* en Afrique montre qu'en plus de L10, l'Afrique centrale abrite également une diversité relativement importante de souches L5.

Malgré la rareté de L10, sa position phylogénétique spécifique et sa présence en Afrique centrale apportent de nouveaux éléments à l'histoire évolutive complexe de *M. africanum*. Le scénario le plus probable actuellement favorise l'Afrique de l'Ouest comme lieu d'origine de tous les variants *M. africanum* [60]. Ce scénario implique que les ancêtres de L5 et L6 ont migré d'Afrique de l'Est et se sont diversifiés en Afrique de l'Ouest, tandis que L9 est retourné en Afrique de l'Est. La découverte de L10 en Afrique centrale avec une ramification intermédiaire entre L5 et L6/L9 peut s'inscrire dans ce scénario mais suggère une migration indépendante de l'Afrique de l'Ouest vers l'Afrique centrale. Alternativement, *M. africanum* pourrait avoir émergé près de l'Afrique centrale, puis migré vers l'ouest et l'est. Ce scénario alternatif nécessiterait toutefois un échantillonnage plus important dans les régions centrales de l'Afrique pour obtenir un soutien réel.

4.3.4 Conclusion

Grâce à l'exploration approfondie des bases de données de séquençage génomique complet et de génotypage, nous avons identifié une lignée rare du complexe *M. tuberculosis*, désignée L10 (proposition), présente en Afrique centrale. Cette lignée se caractérise par une nouvelle région de délétion, des insertions IS*6110*, et 243 SNP, y compris gyrA G7901T, recN C1920096T et dnaG C2621730T. L10 représente un clade frère de L6, principalement présent en Afrique de l'Ouest, et de L9, spécifiquement en Afrique de l'Est, révélant ainsi une pièce potentiellement manquante dans l'histoire évolutive et les migrations de *M. africanum*. Nos découvertes élargissent de ce fait la diversité connue de *M. africanum* en Afrique.

Chapitre 5

Conclusion

5.1 Conclusion générale

L'objectif de ce travail était de proposer une approche globale et à grande échelle pour l'étude et l'exploitation de la donnée génomique brute disponible sur le MTBC, et ce dans tout son ensemble et sa finesse. Avec plus de 160 000 séquences de génomes disponibles publiquement et collectées sur la plateforme du National Center for Biotechnology Information, cette source de données d'une grande richesse restait à analyser. Nous avons constaté l'absence de standardisation des pipelines d'analyse et le manque de bases de données pour partager et étudier les données analysées à un niveau mondial. Les outils ne sont pas, par construction et raison d'être, spécifiques à *M. tuberculosis*, ils gagnent donc en généralité ce qu'ils perdent en spécificité. Nous avons souhaité conserver cette spécificité, tout en atteignant une nouvelle échelle. Un tel outil, sur une quantité massive de données n'était pas encore disponible pour les génomes du MTBC.

Dans un premier temps, il s'agissait de répertorier toute la connaissance sur les caractéristiques du génome de la tuberculose que l'on peut trouver dans la littérature scientifique. Le développement de TB-Annotator a en effet nécessité de regrouper toutes les connaissances existantes sur les différents marqueurs génomiques. Les RD connues ont été collectées manuellement à partir de plus de 20 études, puis organisées et vérifiées. Les SNPs, les *spacers* du CRISPR, les séquences d'IS, et toutes les annotations publiquement disponibles ont été analysées pour leur intégration dans la plateforme. Ces annotations couvrent au final des informations génomiques, géographiques, phénotypiques, phylogénétiques sur des dizaines de milliers de souches.

Dans un deuxième temps, le pipeline de TB-Annotator a été développé. Il a permis d'analyser l'intégralité des séquence publiques du MTBC mais aussi de près de plus de 700 souches privées. Cette analyse a été réalisée en prenant en compte tous les marqueurs spécifiques principaux du MTBC. Le pipeline possède une structure originale où chaque étape de calcul est réutilisée au maximum afin d'augmenter les performances de l'analyse. La réalisation du pipeline a nécessité la conception de plusieurs nouveaux algorithmes et méthodes, notamment pour la détection de variations structurales, méthodes qui pourront être appliquées à d'autres espèces. Les RDs existants et nouveaux, les ISs, l'absence de gènes sont ainsi détectés avec une méthode unique et précise. Les performances du pipeline ont permis de réaliser les calculs sur plus de 160 000 séquences pour obtenir après filtrage plus de 125 000 séquences analysées, avec tous les marqueurs spécifiques extraits et annotés. Le pipeline a démontré ses capacités à grande échelle, et sa capacité d'uniformisation de la donnée.

Une fois le pipeline développé et exécuté, nous avons obtenu un ensemble de données à la même échelle massive que celle du NCBI mais avec les marqueurs génomiques propres au MTBC et entièrement annotés avec la connaissance actuelle présente dans la littérature. Ces données sont difficilement exploitables en l'état et son peu informatives dans leur globalité sans outils adaptés. Comme rappelé en introduction, bien souvent la question

de l'accessibilité des outils est ignorée. Ces données sont amenées à être exploitées aussi bien par des bioinformaticiens que par des biologistes, et ce pour des études allant de la phylogénie à la résistance aux médicaments. C'est pour cela que dans un troisième temps, il a été développé la plateforme d'analyse TB-Annotator. Pour supporter le volume et la complexité des données, nous avons utilisé une méthode originale où la donnée est modélisée de la même manière que dans les moteurs de recherche textuels classiques. Les opérations de recherche textuelles étant peu différentes des opérations que l'on souhaite effectuer sur les données génomiques. Divers calculs ont été développés sur cette base, comme le calcul de l'exclusivité qui permet de déterminer rapidement si un marqueur est propre à un ensemble de souches, ou la recherche de souche similaire grace aux marqueurs remarquables. Les calculs plus classiques comme le calcul de la matrice des distances peuvent aussi être effectués. Toutes ces opérations sont executées en temps réel, c'est à dire que le résultat arrive seulement quelques secondes après le début du calcul et ce sur un interface web facilement accessible. La plateforme permet aussi l'utilisation d'outils externe. Il est ainsi possible d'exporter des fichiers d'alignement, de réaliser des arbres phylogénétiques hors de la plateforme, et de les importer à nouveau dans la plateforme pour pour les explorer avec toute la richesse des données présentes dans TB-Annotator.

Enfin, nous avons montré l'apport de la plateforme avec 3 études dont les découvertes ont été permises par TB-Annotator.

Dans une première étude, TB-Annotator a été utilisé pour l'analyse inaugurale d'un ensemble de 3 000 séquences extraites aléatoirement du NCBI, qui a abouti à la découverte d'une sous-lignée *Asian Ancestral 5* (AAnc5) de la lignée 2. Une analyse ciblée des marqueurs a permis de catégoriser 190 souches comme appartenant au groupe de la lignée 2 (L2) dite *Beijing*. Un examen plus approfondi nous a permis d'identifier les différentes sous-catégories au sein de cette lignée : le *Proto-Beijing*, ainsi que diverses sous-lignées L2 *Ancestral* et *Modern*. Ces groupes étaient tous identifiables par des SNPs cités dans la littérature existante, à l'exception d'un petit clade distinct qui ne présentait aucune mutation indicative de sous-lignées L2 connues. Ce nouveau groupe a été désigné comme *Asian Ancestral 5* (AAnc5).

Une seconde étude a exploré différentes zones d'ombre de la phylogénie, en améliorant la définition des sous-lignées de la lignée 4 et en proposant de nouveaux marqueurs phylogénétiques pour la lignée 4, la lignée 6 et la lignée 7. La lignée 4, bien que fortement séquencée, est hétérogène, avec certaines sous-lignées fortement représentées et subdivisées (par exemple 4.1) et d'autres mal caractérisées (4.5 et 4.7). Cette étude a permis la définition plus précise des sous-lignées 4.7. Enfin, les capacités de calcul de TB-Annotator sur l'intégralité des génomes disponibles ont permis d'améliorer la connaissance des marqueurs sur les lignées 6 et 7 en proposant des SNPs plus exclusifs et dans des régions où le mapping est facilité.

Dans une troisième étude, TB-Annotator a permis de réaliser une découverte phylogénétique importante en identifiant une nouvelle sous-lignée L10 présente en Afrique centrale à partir de 2 souches sur un ensemble de plus de 125 000 souches disponibles dans la plateforme. Pour obtenir ce résultat, nous avons recherché dans la plateforme toutes les souches n'ayant pas de SNPs connus. À ces souches, il a été ajouté des souches représentatives de la diversité du MTBC. Un arbre a été ensuite construit à partir de ces souches, et a révélé toutes les lignées connues et deux souches clairement séparées du reste, que nous avons ensuite identifiées comme la nouvelle L10. L10 représente un clade frère de L6, principalement trouvé en Afrique de l'Ouest, et de L9, spécifiquement en Afrique de l'Est, et a rélévé une pièce potentiellement manquante dans l'histoire évolutive et les migrations de M. africanum.

Enfin, les travaux de cette thèse ont aussi permis la rédaction de deux chapitres d'ouvrages, l'utilisation d'une méthode originale pour l'étude sur la diversité du CRISPR, ainsi que plusieurs autres articles dont les résultats ont été obtenus par analyse avec la plateforme TB-Annotator.

5.2 Le futur de TB-Annotator

De par sa nature, TB-Annotator est en constante évolution. Chaque mois, des centaines de nouvelles séquences sont déposées sur les bases de données publiques. Le pipeline et l'indexation étant totalement automatisés, ces nouvelles souches peuvent continuer à être intégrées pour les années à venir. Actuellement cette indexation des nouvelles séquences est réalisée manuellement, une amélioration future pourrait être d'automatiser ces ajouts, avec une recherche dans les bases de données publiques à intervalle régulier.

L'ajout de nouvelles souches concerne aussi les séquences non publiques. Actuellement cette intégration est réalisée manuellement, avec l'attribution manuelle d'un numéro de séquence pour chaque nouvelle souche. Une amélioration, sur laquelle nous travaillons actuellement, est de rendre possible l'ajout de souches directement depuis la plateforme. Ainsi, chacun pourra ajouter une souche pour analyse et visualiser les résultats sur la plateforme. Cette fonctionnalité n'est pas triviale à implémenter car elle soulève de nombreuses questions. Tout d'abord la confidentialité des données ajoutées, il est nécessaire de mettre en place un mécanisme afin de filtrer les souches privées et qu'elles ne soient pas visibles publiquement sur la plateforme. Ensuite, l'ajout de souche peut créer une pollution de la base de données, par exemple avec des données de mauvaise qualité. Si les souches privées ne sont pas mélangées aux souches publiques, il reste néanmoins la question de la charge de calcul des souches de mauvaise qualité. Un problème qui peut être résolu grâce aux filtres existants. Pour finir, il est nécessaire d'ouvrir les serveurs de calcul au public et d'intégrer des mécanismes de gestion de ces données massives et de répartition de la charge de calcul.

Avec l'ajout de nouvelles souches vient aussi l'ajout de nouvelles références. Cette fonctionnalité est déjà implémentée dans TB-Annotator mais encore sous-exploitée. Actuellement *Mycobacterium leprae*, avec la référence `NC_002677.1`, a été ajouté avec environ 300 souches. Il serait aussi intéressant d'ajouter d'autres références de *Mycobacterium tuberculosis*, mais de par le volume de données, cette opération est longue et couteuse.

TB-Annotator possède la capacité de recalculer partiellement les données. Cette fonctionnalité est utile pour, par exemple, améliorer les algorithmes utilisés. On peut imaginer plusieurs axes d'amélioration des algorithmes. Par exemple, une meilleure gestion du CRISPR avec l'affichage des fragments en plus du spoligotype. Il est aussi possible d'améliorer les capacités de recherche avec de nouveaux calculs, par exemple pour rechercher toutes les souches épidémiologiquement liées à une souche. Avec un développement du séquençage plus important cela permettrait de réaliser un suivi épidémiologique automatisé. Enfin, les données actuellement présentes du TB-Annotator sont uniquement des données génomiques. La richesse de la littérature scientifique sur le MTBC reste encore inexploitée. Dans la section suivante, nous proposons des perspectives sur l'exploitation de ces données ainsi que des méthodes pour la mettre en œuvre.

5.3 Perspectives sur l'unification de la donnée génomique et de la connaissance scientifique associée à l'aide de recherche sémantique et de modèles génératifs

La plateforme TB-Annotator a permis d'unifier et centraliser l'intégralité de la connaissance génomique disponible sur le MTBC. La donnée génomique est analysée pour déterminer des marqueurs génétiques bruts, puis elle est enrichie de manière statique. Cet enrichissement de l'information est dit statique car pour chaque marqueur génétique, les associations possibles à des métadonnées connues ont été manuellement préparées. Si l'on prend l'exemple de la résistance aux médicaments, cette résistance est fréquemment associée à un marqueur comme les SNPs. Cette association est décrite dans toute la richesse de la publication scientifique sur le sujet. On a donc manuellement extrait cette connaissance afin de déterminer l'ensemble des SNPs qui se trouvent être des marqueurs de résistance,

et cette information est ensuite utilisée pour l'annotation des données génomiques. On peut ainsi déterminer si une souche donnée présente une résistance aux médicaments.

La résistance aux médicaments est très étudiée et de nombreuses listes faisant l'association avec des marqueurs génomiques existent (TBDB [223] et WHO [302]). Si l'on prend par contre la résistance phénotypique d'une souche, cette donnée est souvent absente et quand elle est présente, elle l'est uniquement dans les articles associés aux données de séquençage publiées. Lorsqu'un BioProjet est publié sur le NCBI, il contient un ensemble de données de séquençage ainsi que, bien souvent, un article accompagnant ces données. Si lors du BioProjet, la résistance phénotypique des souches a été déterminée, l'information se trouvera bien souvent de manière non structurée dans l'article ou dans les fichiers supplémentaires. Le NCBI permet bien d'annoter les séquences publiées, mais ces annotations sont souvent manquantes. La résistance phénotypique est seulement un exemple du type de données que l'on peut retrouver dans les articles publiés, on y trouve aussi parfois le résultat d'analyses réalisées *in vitro*, pour déterminer par exemple le CRISPR.

On constate que l'extraction de la mine d'information contenue dans les publications scientifiques permettrait de créer sur TB-Annotator le lien entre l'information génomique brute, les marqueurs génétiques déterminés *in silico*, les caractéristiques phénotypiques et les caractéristiques génomiques déterminées *in vitro*. L'information *in vitro* viendrait alors aider à tester et améliorer les algorithmes *in silico* tout en enrichissant la base de données de nouvelles métadonnées. Et les caractéristiques phénotypiques permettraient de découvrir de nouvelles associations entre marqueurs génétiques déterminés *in silico* et leur impact phénotypique.

Les BioProjets associant déjà les articles aux séquences, pour une séquence donnée il est possible de facilement déterminer où est l'article associé et où sont ces informations qui pourraient être extraites pour servir de métadonnées. Cette approche nécessite tout de même d'analyser des milliers d'articles, se pose alors la problématique de cibler les articles susceptibles de contenir des métadonnées pertinentes. Et, de manière plus générale, dans la masse d'information dont nous disposons, rédigée dans un langage avec un vocabulaire spécifique, il n'est pas trivial de trouver une information pertinente ou d'associer l'information issue de multiples articles.

Ce travail d'extraction de l'information contenue dans la publication scientifique est possible manuellement, et c'est ce qui a été réalisé pour le développement de TB-Annotator. C'est une extraction très consommatrice de temps, souvent incomplète, et il est nécessaire de la réaliser en continu, la publication scientifique étant en évolution permanente. Ce travail permet cependant d'obtenir des résultats très fiables qui sont plus difficiles à obtenir avec des méthodes automatisées.

L'avènement des IA génératives ouvre la voie à de nouvelles méthodes d'exploitation de la donnée textuelle. L'extraction d'informations issues d'un texte non structuré devient réalisable à une large échelle et avec une fiabilité jamais atteinte jusqu'à maintenant. Dans cette section, nous explorons la possibilité d'exploiter ces technologies afin d'unifier la donnée génomique de TB-Annotator avec la richesse de la publication scientifique sur le MTBC.

IA générative et texte non structuré

L'intelligence artificielle générative fait référence à une classe d'algorithmes qui apprennent, à partir d'un large corpus de données, à créer de nouveaux contenus sous diverses formes, y compris du texte, des images, des vidéos et de l'audio. Ces algorithmes ont démontré leur capacité à répondre à des requêtes complexes et à accomplir diverses tâches telles que la rédaction d'essais, la réalisation de revues de littérature, le résumé, l'extraction d'informations, etc. La performance de ces large language models (LLMs) dépend de la nature du problème, de l'architecture du modèle et de la qualité des données sur lesquelles les algorithmes sont entraînés.

Parmi ces modèles, on retrouve les modèles Générative Pre-trained Transformers (GPTs)

fortement popularisés par les avancées de OpenAI après le lancement de ChatGPT en novembre 2022. ChatGPT est un agent conversationnel utilisant l'intelligence artificielle générative pour effectuer de nombreuses tâches. Cette implémentation est particulièrement intéressante, car elle offre un niveau de performance très élevé par rapport aux autres modèles disponibles, et ce sur une très large variété de tâches [36].

On peut noter deux composantes essentielles des modèles GPT. La première est l'architecture transformer, qui utilise le mécanisme d'attention, popularisée par la publication par les équipes de Google d'un article désormais bien connu de 2017, intitulé «Attention Is All You Need» [289]. L'attention permet au LLM d'accéder de manière sélective à l'ensemble de l'information des données textuelles en entrée, en appliquant différents degrés d'importance, ou poids, aux informations. La seconde composante concerne la méthode d'entraînement. Jusqu'en 2017, les modèles les plus performants utilisaient principalement l'apprentissage supervisé à partir de données étiquetées manuellement. Il y avait donc un travail très conséquent de la préparation de la donnée et la quantité de données était limitée. En 2018, OpenAI introduit le premier modèle GPT, où un apprentissage autosupervisé est utilisé. L'apprentissage commence par une étape de *pretraining* où le modèle est simplement entrainé à prédire le token suivant dans un paragraphe (un token étant un fragment de mot). Cet entraînement, qui ne nécessite pas de données annotées, permet d'utiliser une masse d'information disponible publiquement, où le modèle apprend à prédire le token suivant sur un grand nombre de textes. Le modèle accumule alors des connaissances basiques comme la structure des phrases pour une langue donnée, mais aussi des connaissances générales sur le monde [219]. Après cet entrainement vient une étape dite de *finetuning*, où le modèle est ajusté pour une tâche donnée. Par exemple, un apprentissage supervisé peut être utilisé pour obtenir des réponses selon un format ou un style particulier. On peut aussi utiliser un apprentissage par renforcement pour rendre le modèle plus aligné avec les attentes d'un utilisateur [207].

Au fil des évolutions, l'IA générative a atteint des capacités générales permettant d'effectuer toute sorte de täches à partir de données textuelles non structurées, et notamment de structurer la donnée textuelle. Par l'apprentissage de formats de sérialisation comme le JSON, les LLM génératifs peuvent générer des structures de données simplement désérialisables. Le LLM, avec des capacités générales, personnalise sont comportement à l'aide de «prompts», c'est-à-dire d'instructions qui sont données avec le texte en entrée du modèle. Le comportement peut aussi être personnalisé en forçant certains tokens dans la réponse, pour par exemple garantir une donnée structurée valide. Un LLM tel que ChatGPT est capable, à partir de paragraphes d'une publication scientifique, de ses annexes et d'un prompt adapté, d'extraire sous forme structurée toutes les métadonnées concernant une souche.

Ce qui était auparavant un problème difficile devient trivial. Cependant la fonctionnalité d'IA générative agit en étape finale. À partir d'un texte relativement court, on obtient une donnée structurée. Dans notre cas les publications sont constituées de dizaines de pages, et les publications se comptent en dizaines de milliers. Le challenge n'est donc pas dans l'extraction de l'information mais dans sa recherche dans un corpus de grande taille. Là aussi l'IA générative a permis de grandes avancées. La recherche de la donnée permet aussi d'autres applications. Par exemple, l'approche appelée Retrieval Augmented Generation (RAG) combine une recherche de texte et une analyse réalisée avec un LLM, ce qui permet à ce dernier d'utiliser des connaissances déterminées et factuelles, et non ses capacités de réponse générales.

IA générative et recherche sémantique

La recherche sémantique désigne une recherche basée sur le sens des mots, par opposition à la recherche lexicale où le moteur de recherche cherche des correspondances littérales des mots de la requête ou de leurs variantes, sans compréhension du sens global de la requête. Via les approches *deep learning*, les modèles peuvent apprendre des repré-

sentations de requêtes et de documents à partir de données étiquetées, où les requêtes et les documents sont transformés en vecteurs de faible dimension (appelés vecteurs denses ou embeddings) dans l'espace de représentation latent. De cette manière, la pertinence peut être mesurée en fonction de la similarité sémantique entre les vecteurs denses.

Le principe du *dense retrieval* est de modéliser l'interaction sémantique entre les requêtes et les textes en se basant sur les représentations apprises dans l'espace sémantique latent. Il existe deux architectures principales pour la récupération dense, le cross-encodeur et le dual-encodeur.

Cross-encodeur Le cross-encodeur considère une paire requête-texte comme une phrase entière. Le texte en entrée du cross-encodeur est la concaténation d'une requête et d'un texte, séparés par un symbole spécial [SEP]. Ensuite, la séquence requête-texte est introduite dans le cross encodeur pour modéliser l'interaction sémantique entre n'importe quels deux tokens de la séquence d'entrée. Il existe différentes stratégies pour obtenir une représentation unique sous forme de vecteur de la séquence requête-texte. En sortie du cross-encodeur on peut placer un classifieur pour obtenir un score entre 0 et 1, qui correspond à une logique apprise par le modèle. Par exemple, un score entre 0 et 1 pour déterminer si le texte répond bien à la requête. Nous avons donc un modèle qui, à partir d'une paire requête-texte, nous permet d'obtenir un score de pertinence.

Prenons l'exemple d'une question utilisateur et d'une base de données contenant N fragments de texte pouvant être une réponse ou non à la question. Nous pouvons créer ainsi N paires de textes pour obtenir le score de chaque paire, ces scores permettent ensuite de trier les fragments de texte et d'obtenir le fragment le plus pertinent. On constate une limitation des cross-encodeurs, pour N très grand le temps de calcul serait très important, car chaque requête impliquerait de comparer toutes les paires de textes. Cependant, il est reconnu que le cross-encodeur est capable d'apprendre des informations d'interaction sémantique plus fines pour une paire requête-texte. Et il est largement reconnu que le cross-encodeur est plus performant pour déterminer la pertinence d'une paire requête-texte [309].

Dual-encodeur Les dual-encodeurs adoptent une approche différente avec l'utilisation de deux modèles. Il existe de nombreuses variations de cette architecture mais nous aborderons la plus simple. Un dual-encodeur basique apprend d'abord des représentations sémantiques pour la requête et le texte avec deux encodeurs séparés, *query embedding* et *text embedding*, au lieu d'un seul modèle comme pour le cross-encodeur. Le score entre les deux représentations est ensuite calculé en utilisant une fonction (par exemple, la similarité cosinus ou le produit scalaire). On constate que les dual-encodeurs sont beaucoup plus efficients et flexibles. Il est possible de recalculer les représentations des textes, ainsi seule une opération (l'embedding de la requête) est nécessaire à chaque recherche. Ensuite un indexe spécialisé, comme le hierarchical navigable small world (HNSW), vient exécuter la fonction de similarité sur toutes les paires de manières approximative et optimisée. Un tradeoff est ainsi effectué entre la précision et la vitesse de calcul.

Recherche en plusieurs phases Les architectures cross-encodeur et dual-encodeur ont respectivement des avantages et des inconvénients. Dans la pratique ces deux architectures sont souvent utilisées conjointement dans les systèmes de recherche sémantique. Le dual-encodeur est utilisé dans une première phase pour filtrer les résultats à grande échelle, et le cross-encodeur vient ensuite trier les résultats. Le cross-encodeur est alors souvent l'implémentation d'un «reranker». Lors du training, le cross-encodeur peut aussi être utilisé pour améliorer le dual-encodeur, c'est une forme de «distillation» de la connaissance capturée par le cross-encodeur.

La recherche lexicale est aussi complémentaire, les modèles sémantiques ayant des difficultés pour recherche des termes précis, car seul le sens de la requête et des textes est

capturé par le modèle.

Entraînement de modèles sémantiques L'entraînement du cross-encodeur et du dual-encodeur est l'étape la plus délicate car elle nécessite des données pour effectuer l'apprentissage, et distinguer les passages pertinents et non pertinents pour une même question. Bien souvent, les données de qualité sont manquantes, particulièrement dans les domaines où le volume de recherche est faible. Il n'est par exemple pas possible d'améliorer les performances du modèle en s'aidant des actions de l'utilisateur comme les liens les plus cliqués dans la page de résultat ou des sondages pour évaluer positivement ou non un résultat.

On peut distinguer deux phases dans l'apprentissage. Une phase de «pretrain» du modèle et une phase de «fine-tune», on appelle ce paradigme «pretrain then fine-tune». Ce principe consiste à entraîner un modèle sur une grande quantité de textes non annoté, puis dans une seconde phase, entraîner le modèle sur une tâche spécifique. Le «pretrain» s'effectue sur des données textuelles générales à grande échelle, le modèle peut ainsi encoder de grandes quantités de connaissances sémantiques, ce qui lui confère une capacité améliorée à représenter la sémantique du contenu textuel. Par exemple, le training de BERT est selon l'objectif «Masked Language Modeling». Sur un grand nombre de phrases issues de sources publiques, un token est remplacé par le token [MASK]. L'objectif est alors que la distribution de probabilité en sortie du modèle maximise la probabilité de prédiction du token masqué. Pour effectuer un tel entraînement, aucune annotation n'est nécessaire, l'entraînement peut être donc effectué à très grande échelle.

La seconde phase, de «fine-tune», est plus délicate car elle s'effectue dans le domaine sémantique de la tâche et nécessite des données adaptées à la tâche à réaliser. Il est coûteux d'acquérir des données d'entraînement à grande échelle pour des questions-réponses dans un domaine précis. MSMARCO et Natural Questions [155] sont les deux plus grands ensembles de données de training pour les questions-réponses générales. Ils ont été créés à partir de moteurs de recherche commerciaux et contiennent respectivement 516 000 et 300 000 questions annotées. Cela reste insuffisant pour couvrir tous les sujets des questions posées par les utilisateurs aux moteurs de recherche [215].

Lors de la création d'un ensemble de données d'entraînement, il est nécessaire de constituer des paires de «positifs» et des paires de «négatifs». Il s'agit de paires question-texte où le texte est une réponse à la question, et de paires question-texte où le texte n'est pas une réponse correcte à la question. L'objectif est bien entendu que le modèle apprenne à associer les textes pertinents aux questions avec des exemples de textes pertinents et des exemples de textes non pertinents. La sélection des négatifs est une étape fondamentale lors du training et a une grande influence sur la qualité des résultats. La quantité de négatifs et la qualité des négatifs sont deux facteurs clés. Par exemple, les «hard negatives» sont des négatifs sémantiquement très proches mais qui ne sont pas pertinents, ils sont donc particulièrement informatifs pour l'apprentissage de la tâche par le modèle. L'entraînement des dual-encodeurs étant nettement plus indirecte que celui des cross-encodeurs, de par la présence de deux modèles, l'apprentissage est d'autant plus difficile. On peut donc utiliser la «distillation» de la connaissance capturée par le cross-encodeur, pour améliorer les données d'entraînement du dual-encodeur.

RocketQA [215] est un exemple d'approche combinant ces différentes techniques dans un pipeline d'entraînement avancé. Cette méthode utilise une organisation lors du training pour augmenter le nombre de négatifs de manière efficiente pour la mémoire («cross-batch negatives»). D'autre part, RocketQA met en place une stratégie afin de maximiser les *hard negatives*. Il est très facile de trouver des négatifs pour une question, par exemple en prenant les positifs d'autres questions, il est beaucoup moins facile de trouver des *hard negatives*. Une approche est de prendre comme négatifs les meilleurs passages déterminés par un autre retriver comme un lexical par exemple. Mais le risque avec cette méthode est de sélectionner des faux négatifs, et ainsi réduire les performances du modèle. Ici,

RocketQA utilise le cross-encodeur, qui est beaucoup plus simple à entrainer avec des données de mauvaise qualité, pour enlever les passages qui pourraient être des faux négatifs. Cette méthode est appelée «Denoised Hard Negatives». Enfin, RocketQA utilise le cross encodeur pour annoter de nouvelles paires non annotées et ainsi augmenter les données d'entrainement. Toutes ces méthodes produisent de bons résultats mais nécessitent tout de même un ensemble de données annotées au départ, ces méthodes permettent donc surtout de tirer parti au maximum des données mais la tâche reste complexe et coûteuse.

LLM pour l'entraînement de modèles sémantiques Les LLM génératifs démontrent des performances inégalées dans une grande variété de tâches. L'arrivée de ces modèles a bouleversé le domaine de la recherche sémantique. En effet, ces modèles peuvent aider à réaliser la tâche la plus complexe lors de l'entraînement de modèles, qui est la constitution des données d'entraînement annotées. Les LLM génératifs sont à la fois coûteux et peu efficients, donc c'est la combinaison avec les architectures classiques telles que décrites précédemment qui permettent d'exploiter leur performances à grande échelle.

On peut distingue deux grandes méthodes exploitants les LLM génératifs : la génération de données synthétiques et la génération de classement. La génération de données synthétiques permet par exemple, avec seulement un corpus de texte, de générer des questions correspondants à des fragments de texte et ainsi produire des paires annotées. Il n'est alors plus nécessaire d'avoir au préalable un ensemble de données de questions. La génération de classement permet à partir d'une question, et d'un petit ensemble de fragments textes, de classer les fragments par ordre de pertinence. Ces deux tâches sont excessivement coûteuses à réaliser en temps normal, mais avec des LLM tels que ChatGPT, le coût est réduit à quelques milliers d'euros. Ces approches reviennent en réalité à distiller la connaissance d'un modèle très large et coûteux dans un modèle efficient comme un dual-encodeur, en générant un ensemble de données annotées.

L'implémentation de la génération de données synthétiques peut se faire de diverses manières. Il est possible de générer des questions à partir de passages pour constituer des données d'entraînement. Mais il est aussi possible de reformuler les questions utilisateur avec le LLM. Les données d'entrées du modèle sont ainsi alignées avec celles vue au training. Les données synthétiques sont générées à l'aide de prompting, c'est-à-dire en trouvant les instructions pour le LLM qui produisent les meilleurs résultats. C'est une étape longue, mais une fois les instructions créées il est possible de générer une quantité massive de données d'entraînement.

Les LLM génératifs offrent des possibilités d'annotation très avancées, qui vont au-delà des annotations que l'on peut trouver sur des données de moteur de recherche commerciaux comme MSMARCO. On peut distinguer 3 catégories d'instuctions [267]. L'approche «pointwise» consiste pour une paire question-passage, à demander au LLM de répondre par oui ou par non si le passage est pertinent. L'approche «pairwise» demande au LLM de classer deux passages pour une question, il s'agit donc de déterminer le passage le plus pertinent. Cela permet de déterminer un classement et un score de pertinence pour chaque paire question-passage. Pour finir, dans l'approche «listwise», on demande au LLM de classer toute une liste de résultats pour une question donnée. Ces approches sont exécutées sur de courtes listes de résultats obtenues à l'aide d'une autre approche, comme une recherche lexicale par exemple.

Conclusion

Les LLM génératifs ouvrent la voie à des modèles de recherche sémantique entrainés dans des domaines spécifiques, à moindre coût, et avec des performances à ce jour inégalées. La combinaison de ces modèles avec le LLM génératif permet de synthétiser et restructurer l'information la plus pertinente sur un sujet donné.

Toutes ces techniques ouvrent ainsi la voie à la création de nouveaux ensembles de métadonnées issus de décennies de publication scientifique. Associer ces données à l'en-

semble des données génomiques de TB-Annotator permettra d'unifier la connaissance sur le MTBC et de servir de source de connaissance pour le suivi et la compréhension du MTBC.

Annexe A

Article du TB-Annotator en anglais

TB-annotator: a scalable web application that allows *in-depth* analysis of very large sets of publicly available *Mycobacterium tuberculosis* complex genomes

Gaetan Senelle, Christophe Guyeux, Guislaine Refrégier,

and Christophe Sola

June 12, 2023

Abstract

Tuberculosis continues to be one of the most threatening bacterial diseases in the world. However, we currently have more than 160,000 Short Read Archives (SRAs) of *Mycobacterium tuberculosis* complex. Such a large amount of data should help to the understanding and the fight against this bacterium. To accomplish this, it would be necessary to thoroughly and comprehensively examine this significant mass of data. This is what TB-Annotator proposes to do, combining a database containing all the diversity of these 160,000 SRAs (at least, SRAs with a reasonable read size and quality), and a fully featured analysis platform to explore and query such a large amount of data. The objective of this article is to present this platform centered on the key notion of exclusivity, to show its numerous capacities (detection of single nucleotide variants, insertion sequences, deletion regions, spoligotyping, etc.) and its general functioning. We will compare TB-Annotator to existing tools for the study of tuberculosis, and show that its objectives are original and have no equivalent at present. The database on which it is based will be presented, with the numerous advanced search queries and screening capacities it offers, and the interest and originality of its phylogenetic tree navigation interface will be detailed. We will end this article with examples of the achievements made possible by the TB-Annotator, followed by avenues for future improvement.

1 Introduction

Tuberculosis remains the most common cause of death from a single infectious bacterium. The World Health Organization's End TB Strategy targets to reduce TB deaths by 95% and to cut new cases by 90% between 2015 and 2035 [1]. An estimated 6.4 million tuberculosis (TB) cases and 1.6 million deaths have occurred in 2021, compared to 5.8 million TB cases and 1.5 million deaths in 2020 worldwide: the eradication of this disease does not seem to be within reach. Tuberculosis has been decimating humanity since antiquity [2]. Its infectious agent was identified by Robert Koch in 1882, and it was later called the *Mycobacterium tuberculosis* bacillus. Since the discovery of Calmette and Guérin's biliary vaccine in the early 20th century and the one of antibiotics shortly thereafter, one of the most promising advances has undoubtedly been the complete sequencing of the bacterium's genome in 1998 [3]. Since then, its sequence has been constantly studied, which has allowed to lift the veil on the complexity of its evolution.

Thanks to genomic studies, it has been shown that tuberculosis in humans is mainly caused by the members of the *Mycobacterium tuberculosis* complex (MTBC). *Mycobacterium tuberculosis sensu stricto* includes 5 lineages: L1 (Indo-Oceanic or EAI), L2 (East-Asian or Beijing), L3 (East-African-Indian or CAS), L4 (Euro-American), and L7 (Ethiopia). *Mycobacterium africanum* includes two other lineages: L5 (West African 1) and L6 (West African 2) [4]. More recently, two new lineages, namely the L8 [5] restricted to the African Great Lakes region and the L9 [6], have been described. Genomic studies have also allowed significant progress to be made in terms of the mechanisms at the origin of its virulence, as well as those explaining the increasing resistance to antibiotics [7,8].

The cost of sequencing has also dropped dramatically over the past 25 years, making the acquisition of new sequences common and affordable. Whole Genome Sequencing (WGS) technologies also referred to as Next Generation Sequencing data (NGS) have made it possible to collect more than 160,000 MTBC genome sequences publicly available on the NCBI SRA database [9], the well known repository for high-throughput sequencing data. Such database provides an essential source of information for the study of the evolution of tuberculosis, but tools are obviously lacking to fully exploit it. The objective of our new platform is not to offer *yet another pipeline method* that does more or less what is already being done but do propose a new paradigm in systems biology of tuberculosis. Its justification comes from the following facts.

First of all, as detailed in Section 3.1, existing tools mainly focus on specific subjects like resistance prediction and genotyping of bacterial isolates. Their aim is not to provide a global and complete approach based on all available SRAs, but something specific on data provided by the user. For instance, some pipelines have been developed to predict potential drug-resistance, like the acknowledged and widely used TB-profiler and/or Phyresse [10,11], Indeed, such pipelines do not have tools to explore and analyse genomic characteristics at a large scale. They usually focus on one type of markers, see Sect. 3.1. They may be poorly equipped with global analytic tools. Some pipelines use a whole bunch of tools, difficult to maintain, difficult to run, and does not allow to scale to the whole MTBC genomes.

As will be seen in Section 3.1, a number of tools are not, by construction and reason to be, specific to *M. tuberculosis*. They therefore gain in generality what they lose in specificity. If they can be applied in a non-specific way to a variety of bacteria, they conversely cannot take advantage of decades of specific knowledge accumulated about the genomic characteristics of the MTBC genomes. In a more subtle way, each pipeline produces different results in incompatible format which makes the construction of a global analysis pipeline, integrating several complementary tools, rather complicated to achieve. In the same vein, it is frequently difficult to compare private strains with already publicly available genomes. To sum up, to set up fully integrated workflows from scratch, that allow an accurate and meaningful analysis of NGS data from clinical MTBC strains, still requires programming expertise, and trained bioinformatics staff. This constrains the application of MTBC NGS analysis to specialized laboratories, leads to a large diversity of analysis pipelines with group-specific solutions that seriously complicates standardized comparison of results.

This explains why we developed the *TB-Annotator* database, that contains informations on raw phylogenetical markers and annotations. It implements information retrieval methods used for text datasets, which have been proven useful for large-scale corpora. This database is at the same massive scale as the SRA one, and it integrates most of known genomic characteristics of MTBC genomes (lineage markers definitions, regions of difference, insertion sequences...). As such, it is the largest pre-analyzed database specific to the *Mycobacterium tuberculosis* complex. Moreover, this database comes with an ergonomic and original interface, allowing to make complex and advanced

queries or to create complex filters, and to investigate with a yet unachieved depth, any specific phylogenetic tree of your choice. As will be seen below, TB-annotator has already allowed us to revisit a certain amount of knowledge on two lineages of tuberculosis [12, 13]. While access is currently restricted (interested readers can apply to the authors for access), it is intended to become public soon.

To the best of our knowledge, such a platform that would process massive amount of data is not yet available for *Mycobacterium tuberculosis* complex genomes. Concretely, this large amount of still unexploited spatio-temporal genomic diversity data could constitute in a near future a gold-mine of a yet hidden knowledge, that would allow to perform a kind of archeological dive into the past of specific infectious diseases history. With more than 160,000 SRA available, and constantly updated, the field is quite mature, and part of the work will consist in being able to discriminate what are the information that will really improve our global understanding of the tuberculosis pandemia, from the data that will be either useless or contribute to the well-known overlearning process. By doing so, one of the final aim of this evolving project could be to design for tuberculosis a similar project than the one that was launched just before the Covid-19 sanitary crisis, whose ambition was to perform real-time tracking of pathogens [14], or even to more ambitiously create a platform that would allow to perform deep-learning for an improved characterization, diagnostic or treatment of TB diseases applied to patient or communities, in relation to genomic data analysis.

The rest of this article is structured as follows. In the following section, we present TB-Annotator in detail, both in terms of the database and its interface. We discuss its originality in Section 3, showing that its objectives are different from the tools currently used by the community. Examples of its implementation are also given in this section, for illustration purposes. This article ends with a perspective, in which the interest of the TB-annotator is recalled, with new studies that we wish to carry out or are currently in progress.

2 TB-annotator description

TB-Annotator is made up of two main components. First the pipeline which massively analyse all MTBC publicly available genome. And second, the analysis platform to explore and search this data.

2.1 Pipeline architecture

First of all, the pipeline uses Snakemake [15] as the execution engine. Snakemake is designed to build reproductible and scalable data pipeline by leveraging a declarative workflow definition and modern execution platforms such as cloud (for example see Kubernetes [16]). This workflow engine was selected over cloud native workflow engines such as Argo Workflows [17] for its ease of use, ease of deployment on personal desktops, and the facilitated integration with common bioinformatics tools.

The pipeline starts with a list of accession numbers provided by the user. Sequences are then downloaded from Sequence Read Archive (SRA) [18], with fastq-dump, as paired-end or single-end FASTQ files. This first step of the pipeline is throttled to reduce the number of simultaneous connections to the NCBI. FASTQ files contain sequencing reads with the sequence and a quality score for each base, the PHRED score [19]. These scores are used in the reads pre-processing and mapping steps to assess the reads quality, therefore FASTQ is the only supported input format.

For non-publicly available sequences, the pipeline accepts compressed FASTQ files. As a convention, sequences from private sources are given a custom identification number starting with the prefix *CUS* (for example CUS000001). The pipeline is preconfigured

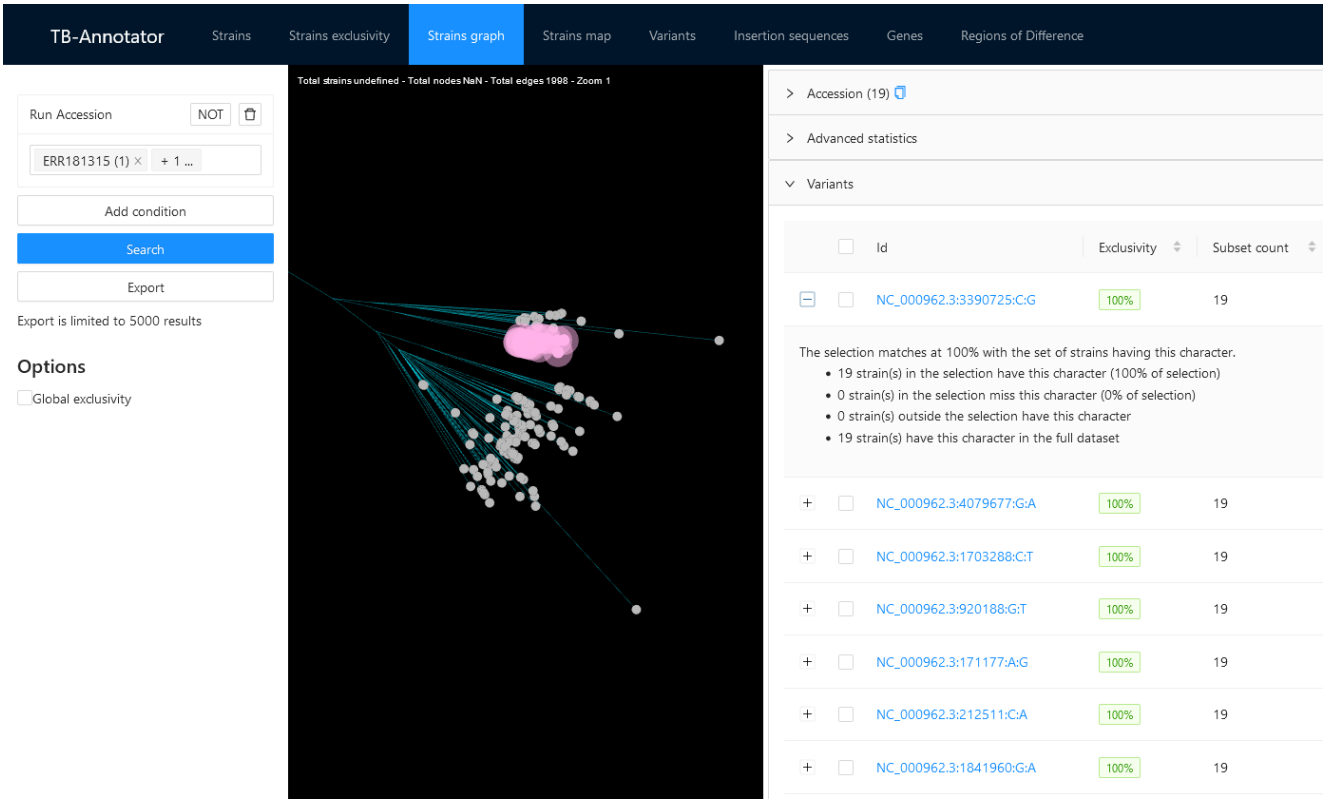


Fig 1. Graph visualization

to detect this naming scheme, and to analyse *CUS* sequences at the same time as SRA sequences.

The pipeline DAG (directed acyclic graph) and computations are organized to maximize data reuse and incremental enhancements. It allows to perform new analyses and computations without doing again costly steps like SRA download.

2.2 Read preparation

Genome coverage is estimated based on the number of reads in the FASTQ file. To ensure the speed of the pipeline remains constant, the FASTQ file is randomly downsampled to a configurable estimated genome coverage using SeqKit [20].

Fastp [21], an all-in-one and high performance tool, is used for quality control on the reads and to pre-process them. In the pre-processing part, the adapters used in library preparation process during sequencing are trimmed. Low quality bases at ends of reads are cut and reads with too many low quality bases are removed. For paired-end reads that are overlapping, they are corrected based on quality score if they are not a perfect reverse-complement. Some advanced corrections are also performed like polyG trimming and UMI preprocessing [22].

Following the read quality control and pre-processing, a report is outputted in html and json formats. The report contains informations and statistics about the different parts of the reads filtered, base contents, and duplication rates. This report can be used to evaluate the sequencing process, and detect problems like contaminated reads, base content biases and over-represented sequences. Original FASTQ files are removed after this process to reduce disk space usage.

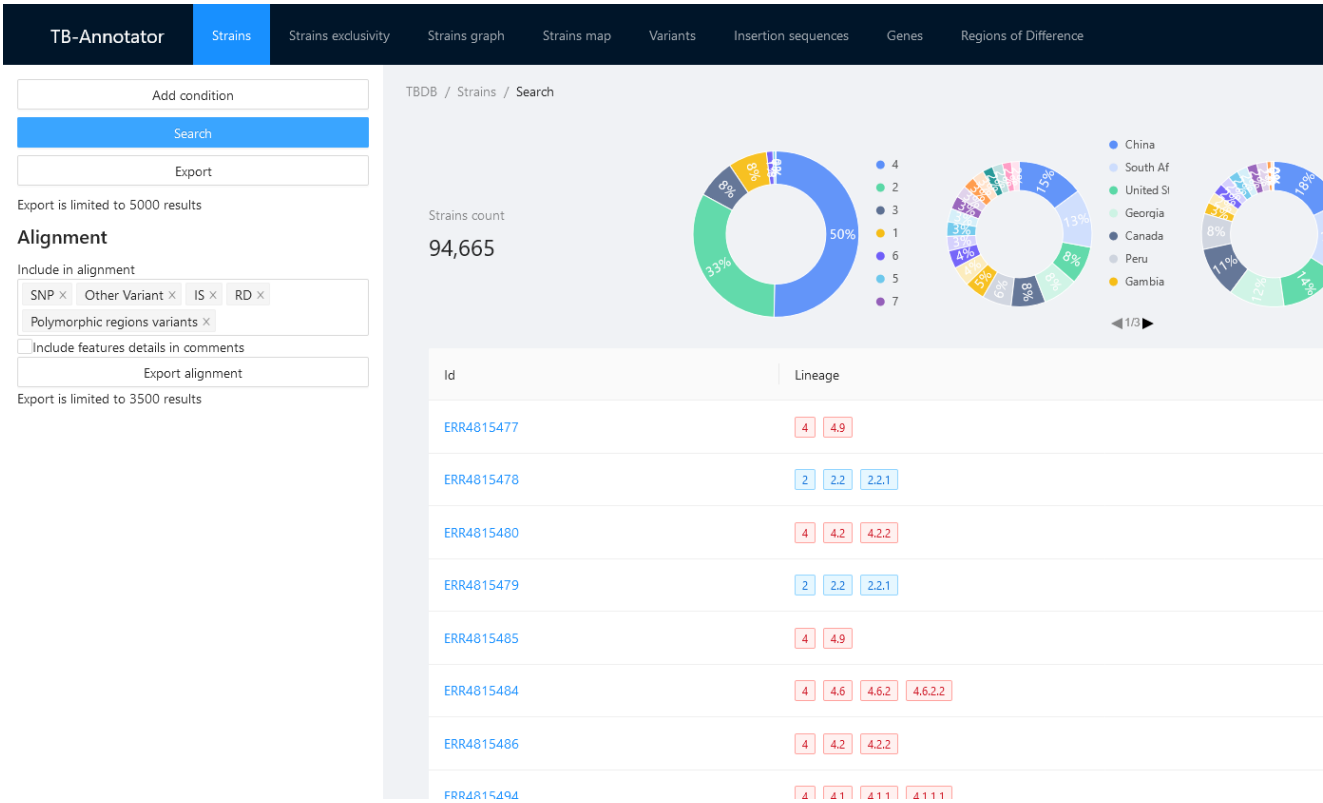


Fig 2. Database overview

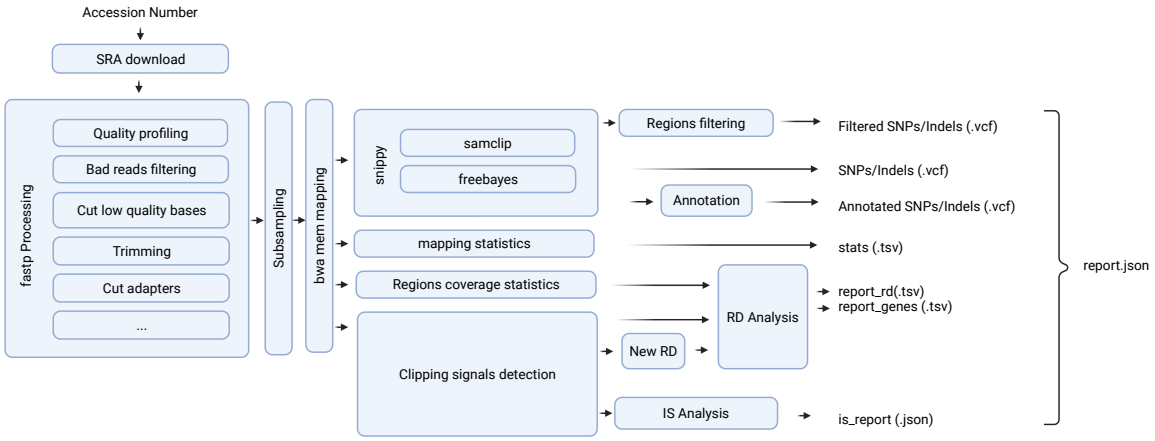


Fig 3. Pipeline overview

2.3 Read mapping

Processed reads are mapped on a reference genome (NC_000962.3 by default) using BWA-MEM [23]. BWA-MEM is one of the fastest mapper [24], while being close to Novoalign in terms of accuracy [25]. Duplicate reads are marked using SAMTOOLS [26] for the next tools in the pipeline, including freebayes, to prevent bias in the variant

calling due to PCR duplicates that might cause the algorithm to identify an error during the amplification as a true variant [27].

The sequence alignment is saved in CRAM format, a compressed format based on the reference sequence [28]. The resulting file is small enough to be kept on hard drive at low cost, and to be able to perform further analysis or to execute an updated pipeline in the future without executing the download and pre-processing step a second time. This optimization is made possible thanks to Snakemake ability to continue partially executed workflows. The computing time is decreased by using the sequence alignment as a basis for all the subsequent steps of the pipeline, instead of executing analysis on raw reads.

Finally statistics are computed with SAMTOOLS using the sequence alignment file, like mean read depth and the proportion of covered bases. This information can be used to assess the quality of the sample and the quality of sequencing. This information will also be used in the rest of the pipeline for statistical analysis.

2.4 Variant calling and annotation

Snippy [29] is used for variant calling, which is one of the top performing pipeline for bacterial genomes [30]. Snippy takes as input the already aligned reads in compressed CRAM format instead of letting it aligns the read (by default with bwa-mem too) to bam format. Local realignment was not performed before variant calling because it had a low impact with the tools used [30].

Snippy starts with the use of Samclip [31] on aligned reads to avoid false SNP calling close to structural variations. It works by removing clipped reads inside contigs. The resulting alignment mainly contains long aligned reads and thus providing higher confidence in variations called. Variants are then called with Freebayes [32]. The final output of this step is a VCF file.

Some regions of the MTBC genomes are known to be difficult to map including mobile elements, Proline-Glutamate (PE)/Proline-Proline-Glutamate (PPE) genes and repetitive regions. These regions usually contain a number of soft-clipped reads that are removed by samclip, but some of them are kept, for example reads corresponding to insertions sequences mapped to IS regions of the MTBC reference. These problematic regions are removed in number of studies [33]. An additional filtered VCF file is outputted in which variants falling in these regions are removed.

Variants are annotated using SnpEff [34], a variant annotation and effect prediction tool. SnpEff annotates and predicts the effects of genetic variants (such as amino acid changes). We identify variants using their SPDI notation [35], this allows to have a globally unique identifier for each variant. The alternate representation in which the Deletion field is a string containing the literal sequence to delete [35] is used to avoid the need for reference sequence when reading the SPDI.

2.5 Read clipping analysis

When a read partially matches the reference sequence, the aligner usually removes the unmatched part, this process is called clipping. Two types of clipping exist: (1) hard clipping where the removed part is indeed removed from the resulting alignment file, and (2) soft clipping where the removed sequence is kept and its position is signaled in the CIGAR string. By default BWA-mem uses soft clipping for primary alignment. A large number of clipped reads at the same positions is usually a sign a of Structural Variant (SV), and it is used for general SV calling [36].

We call *clipping signal* the position where a configurable amount of reads is clipped. A clipping signal can be on the left side or the right side of reads (Figure 4). In the clipping signal detection step, a custom Python script creates the list of clipped reads,

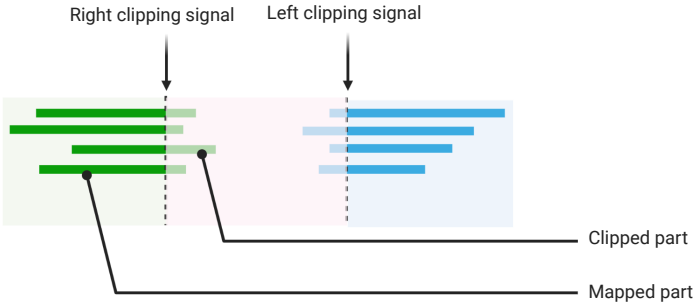


Fig 4. Clipping signals

the position of the left clipping and right clipping and the clipped sequence. The list of clipping signals is created using *bedtools groupby* [37] on the output of the Python script.

This step is only a pre-processing for subsequent steps. The clipping information can be used to detect SV such as deletion and insertions. And contrary to the general case of SV detection, we can use this information to detect known MTBC genome features more easily and more precisely.

2.6 Detection of the present/absence of known RD and genes

Known Regions of Difference (RD) were manually collected from more than 20 studies, curated and checked. The positions were converted to positions on the reference genome H37Rv NC_000962.3. When this position was not available in the original article, it was inferred using various methods like mapping the PCR junction, mapping example strains containing this deletion, or extracting the position from other references like *M.bovis*.

This step focuses on assessing the presence/absence of genes and large RD (RD much smaller than read length, like pks 1/15 7bp insertion, are found by the variant calling step [38]). A read depth histogram is computed on the aligned reads using BEDTOOLS [37]. For each region (RD and genes) multiple statistics are computed like mean, median, min, max coverage and the ratio $\frac{\text{region mean coverage}}{\text{global mean coverage}}$. Most of existing approaches for *in silico* RD detection uses these metrics [6, 39, 40]. With this method, we miss various information, for example: if the region is fully deleted or not, if the region is part of a larger deletion or precisely deleted, if it is probable that the noticed drop in read depth is indeed related to a deletion. To overcome some of these issues, we compute the percentage of very low read depth in the given region using a configurable threshold. The resulting value can be used as an indication of partially deleted gene for example.

The second method is based on clipping signals extracted from a previous step. We use these signals to find what we call *high quality deletions*, which are regions for which we are highly confident that they are precise deletions (and not false positives due to lower read depth in the region, for example). Clipping signals close to the sides of the region are searched using a configurable flank. Clipped part is aligned on the opposite side of the region as described in Figure 5. When a configurable number of sequences precisely aligns on both sides, we mark the region as *high quality*.

Flanks are used when searching clipping signals mainly due to the way reads are aligned. The clipping signal is not always close to the end of deletions. This is the case for example when the sequence before the start of the deletion equals the sequence at the end of the regions. This case is illustrated on Figure 6. Because we know approximately

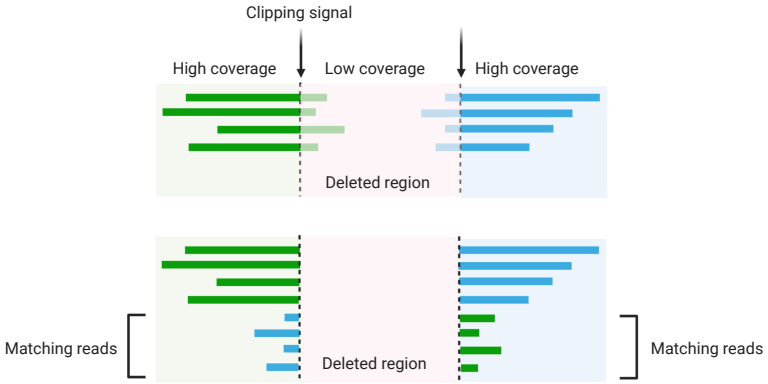


Fig 5. Mapping of clipped sequences

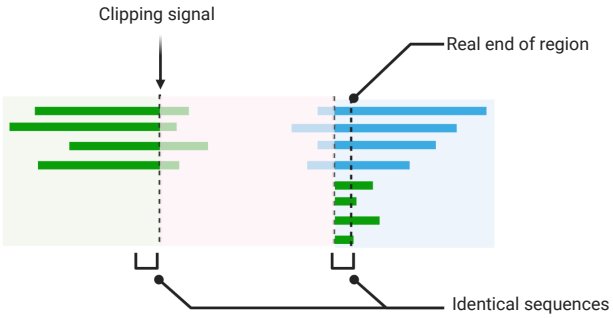


Fig 6. Difference between clipping signal and ends of deleted regions

where the two clipping signals are for a given region, this decreases the risk of mapping clipped sequences at a wrong position in the genome. Not all regions can be considered high quality, like RD falling in larger deletions. These regions are still detected using read depth variation like explained earlier. This case is illustrated in Figure 7.

The pipeline comes with two list of regions, genes and known RDs, while additional regions are configurable with bed files. For each list of regions, a TSV report is created, with all statistics described earlier and a score indicating if the region is considered as *high quality*. At the end of this step, all the searched regions are outputted. The final choice (if a region is deleted or not) is left to the last step by the creation of the final report. This allows to keep detailed information on regions for further analysis.

2.7 Detection of new RD

The pipeline allows the detection of RDs that are not documented in existing studies. To ensure the consistency and precision of positions detected across all strains analysed, this step is also based on previously detected clipping signals. Possible regions of deletion are searched by creating pairs of consecutive clipping signals, and a name is generated for each of these pairs. These regions are then analysed at the same time as known regions to increase pipeline speed.

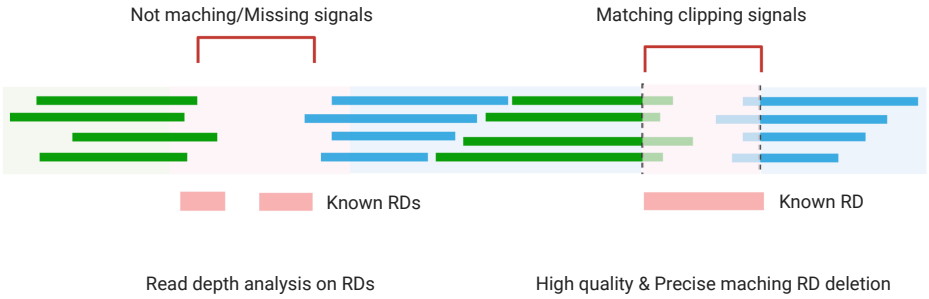


Fig 7. Different methods of detection of regions deletions

2.8 Detection of CRISPR

Reads are mapped using BWA-MEM [23] on the sequences of all known spacers. For each spacer, instead of outputting binary presence of the spacer, we output statistics to help evaluate the probability of the presence or absence of the spacer.

2.9 Detection of insertion sequences

We collected sequences as fasta files from 22 known MTBC insertion sequences (https://www-is.biotoul.fr/).

Clipping signals detected in a previous step are used to detect insertion sequences. Pairs of clipping signals separated by a configurable small distance are searched. For each pair of clipping signals, clipped sequences from the reads are extracted (Figure 8) and mapped using BWA-MEM [23] on a fasta file containing all known insertion sequences.

The matching insertion sequence is determined by analysing the resulting mapping file. We find the first sequence where (1) clipped sequences are mapped on both sides of the reference insertion sequence, (2) left clipped sequences align on the same position, and (3) right clipped sequences align on the same position. This process ensures that the start and the end of the insertion is coherent. This method also allows to detect IS insertion orientation depending on which side of the reference insertion sequence reads are mapped. When the inserted sequence is a subsequence of an IS, we also store the start and the end of the subsequence relatively to the reference IS.

For insertion sequences contained in the reference, RD detection algorithm is used to determine if the insertion sequence is deleted. This detection is made possible thanks to the clipping signal algorithm. Achieving this detection would not be feasible using RD algorithms relying on read depth, as reads from other insertion sequences may map to the deleted region.

2.10 Final report

The output of all the steps of the pipeline are gathered in a single JSON report containing:

- the list of variants found and their annotations,
- the list of missing genes,
- the list of missing region of differences,

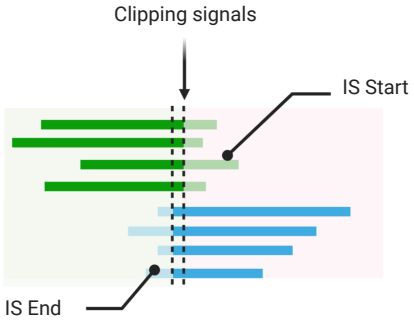


Fig 8. Insertion sequence detection using clipping signals

- coverage statistics for each gene and RD,
- the list of insertion sequences,
- and the quality report from Fastp.

2.11 Pipeline execution

We run the pipeline with 128 threads on a Dell PowerEdge R740 hosted at the Mésocentre de calcul de Franche-Comté. The system is using a 16 core Intel(R) Xeon(R) Gold 6226R CPU @ 2.90GHz with 128GiB of RAM.

MTBC SRA are analysed by batches of 15,000, each batch taking approximately one week to complete. Task scheduling is handled by Snakemake using the greedy scheduler to decrease the job graph computation time.

2.12 Data search

In order to search the analysed strains, all the report data is enriched and indexed in an inverted index. Inverted index was chosen due to the fact that most queries consist of searching strains having a certain character (like a specific variant), and such queries are very efficient using inverted indexes.

An indexing pipeline, developed in C# using Dataflow API, is used to read the genomic pipeline JSON reports, and to index the data in Elasticsearch 8. The report is first read in memory, splitted in 4 types of data (strain, variant, RD, IS), annotated and finally indexed in separate indices. The indexing performance is close to 20 reports per second using a 8 cores AMD Ryzen 7 2700 Processor @ 3.20GHz with 32GiB of RAM, with a local instance of the database.

2.13 Feature exclusivity computation

The feature exclusivity allows to select a subset of clinical isolates (called foreground set) among the set (called background set) of all genome sequences publicly available on the NCBI SRA database, and to determine if the features of those isolates (variants, RD, IS) are exclusive to the subset. TB-Annotator is able to run such analysis on very large amount of strains by leveraging the inverted index global ordinals. Using a custom script, the Jaccard coefficient is computed to be used as a score on every feature of the strains.

2.14 External annotations during indexing

The Snakemake pipeline is only based on genomic data, all additional enrichments are done during the indexing.

Each variant indexed in the database is annotated for resistance based on TBDB [41] and WHO [42] data. When a SNP is a known lineage marker, the SNP is annotated with a reference to the author [43–48]

For each isolate, the indexer starts by collecting all public metadata available in the NCBI database. Results are cached for faster re-indexation and to limit number of request to NCBI API. The isolate is then annotated for resistance based on variants annotation associated to this isolate.

Genes, for their parts, are annotated using NCBI NC_000962.3 metadata.

2.15 Web platform

TB-Annotator Web platform allows to browse, search and run advanced analyses on the database from the user browser. User is able to perform fast queries on most attributes of strains, variants, insertion sequences, genes and RD using the inverted index.

The creation of alignments to be used in tools like RAxML is done by querying the database in batches, and dynamically computing a map between each document and the features it contains. The map is then used to generate an alignment in FASTA format. In case where other features than SNP are exported, a binary alignment is used. The details of columns is also exported at the end of the FASTA file in comments.

For the *strain* graph feature, user phylogenetic trees in Newick are first parsed in the browser. The tree is then rendered using WebGL and equal-angle method layout algorithm also implemented in the browser. The name of each node in the graph is then extracted to be sent to the server to be used as list of accession. It allows the use of feature exclusivity computation against arbitrary phylogenetic trees. The list of all displayed strains is used as a background set for exclusivity comparison (the whole database can also be used as background set).

3 Discussion

The aim of this section is to show that, on the one hand, the TB-annotator is original and complementary to existing tools and, on the other hand, that it is useful, having already enabled us to make new discoveries within the MTBC.

3.1 Comparison with existing tools for *M.tuberculosis* genome analysis

In this section we will compare TB-Annotator to existing web platforms and other databases available.

3.1.1 WGS data analysis tools

TBprofiler [49] is another next-generation sequencing analysis tool for *Mycobacterium tuberculosis* that focuses on detecting drug resistance mutations. TBprofiler also provides a web interface, but its scope is limited to drug resistance analysis. TB-Annotator goes beyond this by allowing analysis and visualization of insertions, difference regions, and missing genes, thus providing a more comprehensive analysis of the genomes.

CASTB (Comprehensive Analysis Server for the Mycobacterium tuberculosis Complex [50]) is a database and web server that allows the analysis of whole-genome sequences of *M. tuberculosis* complex strains. The tool provides information on drug

resistance and lineage identification through a series of analyses, such as SNP-based phylogeny and prediction of drug resistance mutations. However, unlike TB-Annotator, CASTB is not designed to be used for real-time analysis of large-scale genomic data, and it does not have a comprehensive annotation pipeline to detect genomic variations beyond known resistance mutations.

Mykrobe [51] is a suite of tools for the identification of drug resistance in bacterial pathogens, including *M. tuberculosis*. The tool provides predictions for resistance to first-line and second-line drugs, and it also allows the identification of lineage for isolates. However, mykrobe is not designed for comprehensive annotation of genomic variations beyond resistance mutations.

Compared to these tools, TB-Annotator provides a comprehensive pipeline for the detection of genomic variations, including SNPs, insertions, deletions, and large-scale genomic rearrangements. Additionally, TB-Annotator allows for the annotation of genomic variations beyond known drug resistance mutations, enabling researchers to study the evolution and transmission of *M. tuberculosis* strains in greater detail. Furthermore, TB-Annotator provides a user-friendly web platform that allows for real-time analysis and visualization of large-scale genomic data. Overall, TB-Annotator stands out as a unique tool that provides a comprehensive and user-friendly approach to the annotation and analysis of genomic data in the context of *M. tuberculosis* research.

3.1.2 Genetic profiling

MIRU-VNTRplus [52] is an online analysis tool designed for the identification, typing, and phylogenetic comparison of MTBC strains. The tool is based on the typing method called MIRU-VNTR (Mycobacterial Interspersed Repetitive Units-Variable Number Tandem Repeats), which is a PCR-based molecular typing technique. MIRU-VNTRplus provides a platform for comparing new isolates to a reference database, allowing researchers and clinicians to better understand the genetic diversity, epidemiology, and transmission patterns of tuberculosis. This tool is therefore very different from ours: it focuses on a single subject of analysis (MIRU-VNTRs), of exclusively experimental origin, and it handles only one isolate at a time (the one submitted by the user). Our tool does not have as input experimental MIRU-VNTR extraction data, but read files. It is unable to analyse these MIRU-VNTRs, as the size of the reads makes it impossible to count copies of motifs of interest. However, it does produce all possible analyses on sequencing data, and provides a comparative analysis of all available genomes (with sufficient quality and read size).

SITVITWEB [53] is an online database and analysis tool for Spoligotyping-based genotyping of *Mycobacterium tuberculosis* complex (MTBC) strains. Spoligotyping is a PCR-based method that detects the presence or absence of specific spacer sequences in the direct repeat region of the MTBC genome. SITVITWEB allows users to query the database with new Spoligotype patterns, providing information on global distribution, phylogenetic classification, and epidemiological context. The tool aids researchers and public health professionals in understanding the population structure, geographical distribution, and transmission patterns of tuberculosis strains.

3.1.3 Gene annotations and expression profiles

TBDB [54] is an integrated database containing information about *M. tuberculosis* genes, proteins, and drug resistance mutations. It serves as a valuable resource for researchers by consolidating data from various sources and providing tools for data analysis and visualization. Although TBDB offers a rich collection of genomic information, it does not offer an analysis pipeline for whole-genome sequencing data, as TB-Annotator does. TB-Annotator not only analyzes and annotates genomic data but

also provides advanced search capabilities and a platform for running advanced analyses on the resulting data.

TubercuList [55] and MycoBrowser [56] are online databases containing annotated genomic information for *Mycobacterium tuberculosis*. These resources allow for gene and protein exploration, providing detailed information on gene function, gene ontologies, and metabolic pathways. However, they do not provide the analysis, visualization, and search capabilities that are central to TB-Annotator. TB-Annotator's comprehensive approach to whole-genome sequencing data analysis, along with its advanced search and visualization features, sets it apart from these databases.

PATRIC [57] is an extensive online resource for bacterial pathogens, offering tools for genomic and comparative analysis. PATRIC covers a wide range of bacterial species, including *M. tuberculosis*. While it provides valuable data and analysis tools for researchers working on various bacterial pathogens, it is not specifically tailored for *M. tuberculosis* and does not provide the same depth of analysis, visualization, and strain-specific data as TB-Annotator. TB-Annotator is designed exclusively for *M. tuberculosis*, enabling a more focused and comprehensive analysis of this particular organism.

BioCyc [58] is a collection of pathway and genome databases that cover various organisms, including *Mycobacterium tuberculosis*. BioCyc offers valuable information on metabolic pathways, gene function, and genome annotations. However, it does not provide the comprehensive genome analysis, visualization, and search capabilities that TB-Annotator offers. TB-Annotator is designed specifically for *M. tuberculosis*, allowing it to provide a more detailed and focused view of the genomic features of this organism, along with advanced search and analysis options.

3.1.4 Other partly comparable websites

Enterobase [59] is a publicly accessible database of bacterial pathogens, including *Mycobacterium tuberculosis*. It allows users to submit sequencing data and compare it with other isolates to identify outbreak clusters and track the spread of antimicrobial resistance. While it has some similar functionality to TB-Annotator, it does not provide the same level of annotation and analysis of genomic features. Additionally, Enterobase does not currently include all available *M. tuberculosis* sequencing data, whereas TB-Annotator incorporates a large portion of the publicly available SRA data.

PathogenSeq [60] is a web-based platform that provides tools for the analysis of microbial genomic data, including *M. tuberculosis*. It allows users to upload their own data or use pre-loaded datasets and provides tools for assembly, annotation, and comparative genomics. However, unlike TB-Annotator, PathogenSeq is focused more on providing raw data analysis tools rather than pre-annotated data. Additionally, it does not include some of the features of TB-Annotator, such as the ability to search for specific genomic features, investigate phylogenetic trees, and perform feature exclusivity analysis.

Nextstrain [14] is a web-based platform that provides real-time tracking of global infectious disease outbreaks, including M. tuberculosis. It allows users to view phylogenetic trees of viral and bacterial sequences, track the spread of outbreaks, and identify new genetic variants. While it provides some similar functionality to TB-Annotator, it does not provide the same level of annotation and analysis of genomic features, nor does it include some of the search and filtering options of TB-Annotator. Additionally, Nextstrain is focused more on tracking the spread of disease in real-time, whereas TB-Annotator is more focused on providing a comprehensive database of M. tuberculosis genomic features.

CNGBdb [61], finally, is a publicly accessible database of genomic data, including *M. tuberculosis*. It provides tools for searching and analyzing genomic data, as well as data visualization and comparison. While it has some similar functionality to TB-Annotator, it does not provide the same level of annotation and analysis of genomic features, nor

does it include some of the search and filtering options of TB-Annotator. Additionally, CNGBdb does not include all available *M. tuberculosis* sequencing data, whereas TB-Annotator does.

3.1.5 Summary

In conclusion, TB-Annotator is a unique tool that stands out for its ability to provide a comprehensive analysis of *M. tuberculosis* genomes. Unlike most of the existing tools, TB-Annotator integrates several steps that range from quality filtering of reads to annotation and indexing of detected variants, insertion sequences, and regions of differences. Furthermore, TB-Annotator has a web-based interface that allows easy visualization and mining of data from the database. Compared to other tools, TB-Annotator offers several additional features, such as the detection of new RDs, the ability to detect insertion sequences, and the computation of feature exclusivity. Moreover, TB-Annotator provides external annotations during indexing that improve the resistance annotation of indexed variants. Finally, TB-Annotator is capable of analyzing a large number of genomes simultaneously, making it ideal for population-based studies.

Overall, TB-Annotator offers a powerful and user-friendly platform for exploring M. tuberculosis genomic diversity, and its unique features make it a valuable addition to the existing tools. By providing a comprehensive analysis of genomic data, TB-Annotator will facilitate better understanding of TB epidemiology and resistance patterns, which will ultimately help in the design of more effective control strategies. Additionally, its web-based interface and the ability to analyze large datasets will make it a useful resource for researchers and public health professionals alike.

3.2 Examples of using the TB-annotator to answer specific questions

3.2.1 Connection between two historical tuberculosis outbreak sites in Japan, Honshu, by a new ancestral Mycobacterium tuberculosis L2 sublineage [62]

In this study, we describe a historically endemic ancestral sublineage of L2, known as AAnc5, based on samples collected in the Tochigi Prefecture in central Japan. This sublineage is closely related to the Japanese G3 group identified in 2012 and is presumed to be named L2.2.A in a recent review. The findings of this study strengthen the phylogenetic relevance of this sublineage within the global L2 evolutionary history, showing that it was historically transmitted in several Japanese cities.

Tuberculosis is an ancient disease in human history, but its exact emergence in Asia remains uncertain. Early TB outbreaks in Japan may be connected to population migrations between the 5th century BC and the 3rd century AD. Tuberculosis was known to be present in ancient Japan under the name of rôga, which was used in Chinese medicine. Traces of this disease can be found in ancient Chinese medical texts.

A critical aspect of this study is the use of the TB-Annotator pipeline, which has greatly facilitated the analysis of large amounts of data contained in Sequence Read Archive (SRA) files. By analyzing over 50,000 characters, including repeated sequences and single nucleotide polymorphisms (SNPs), TB-Annotator enables a more in-depth examination of the complex historical phylodynamics of all *Mycobacterium tuberculosis* complex (MTBC) lineages. This tool has played a significant role in obtaining the results presented in this study and in mapping the endemic L2 ancestral sublineage from Japan onto the global MTBC phylogeny.

Among the specific characteristics of AAnc5, the authors describe a non-synonymous mutation in the rpoC gene. This mutation is generally found in epidemiologically suc-

cessful isolates that also contain specific rpoB gene mutations. The use of TB-Annotator has been crucial in identifying such unique features within the AAnc5 sublineage.

Estimations of the coalescence of AAnc5 sublineage landmarks range from 280-310 years to 760-800 years before the present. The Tochigi Prefecture is famous for its Ashio copper mine, which began operations in the early 17th century. The Ashio mine could have been a location for AAnc5's expansion and diversification.

In conclusion, the TB-Annotator pipeline has been instrumental in mapping an endemic L2 ancestral sublineage from Japan onto the global MTBC phylogeny and designating it as Asia Ancestral 5 (AAnc5). This sublineage possesses many specific characteristics that distinguish it from other ancestral sublineages described so far in L2. The discovery of AAnc5 and the extensive insights provided by TB-Annotator open new avenues for research into the history of L2 in Southeast Asia.

3.2.2 Tuberculosis in Nigeria: new insights from whole genome sequencing data (submitted work)

In this study, we investigated the genomic diversity of *Mycobacterium tuberculosis* lineages in Nigeria, with a focus on *M. africanum* (L5-L6) distribution. TB-annotator, an essential tool in the analysis, provides a deeper understanding of the genomic diversity and the rarer sub-lineages. The study reveals a fully diversified L5 lineage, while the L6 lineage is poorly represented. The absence of certain sub-lineages raises questions and will require further clarification. The low presence of L5-L6 in Nigeria suggests a possible disappearance of *M. africanum* in the region, a topic that has been debated in recent years.

TB-annotator plays a crucial role in analyzing these lineages, as it helps identify new isolates and their distribution. For example, the study found that the number of known isolates for the 511212 sub-lineage doubled, allowing for a better understanding of its specificity. The tool also aids in understanding the presence of certain sub-lineages in Nigeria, such as one-third of the L5 isolates being found in the country. This information is valuable for further research and could potentially lead to the discovery of new trends and patterns in the distribution of tuberculosis strains.

The study also discusses the importance of monitoring drug-resistance in Nigeria. A decade after the initial work, researchers found that 39.49% of patients were resistant to at least both INH and RIF, leading to an MDR rate well above 40%. This highlights the need for ongoing surveillance and the implementation of effective tuberculosis control programs. TB-annotator proves to be essential in this regard, as it helps assess drug-resistance status and determine the prevalence of resistant strains in specific regions.

In summary, this research investigates the genomic diversity of *Mycobacterium tuberculosis* lineages in Nigeria, emphasizing the significance of TB-annotator in obtaining a comprehensive understanding of these lineages and their drug-resistance profiles. The study highlights the potential disappearance of *M. africanum* in Nigeria and the increasing importance of drug-resistance surveillance in the country. TB-annotator serves as a valuable tool for researchers to uncover new insights, track the spread of specific sub-lineages, and ultimately contribute to the development of more effective tuberculosis control strategies.

3.3 Examples of using the TB-annotator to answer more general questions: the case of the L5

The submitted article entitled "An updated evolutionary history and taxonomy of *Mycobacterium tuberculosis* lineage 5, also called *M. africanum*" focuses on lineage 5 (L5) of the *Mycobacterium tuberculosis* complex, which is a significant cause of tuberculosis

in West and Central Africa. Despite the importance of L5, it has been poorly represented in public databases, which has made it challenging to reconstruct its evolutionary history and taxonomy. The authors aimed to build an exhaustive collection of representative L5 genomes and analyze their structure using TB-Annotator.

The study provides a considerable increase in the phylogenetical analysis of L5, offering a better description of basal groups, the identification of new reliable SNPs at each subbranch, and an integrated phylogenetic reconstruction of the entire L5 lineage. The authors note that diversification, i.e., transmission, was pervasive in the beginning of L5 history but diminished in later evolutionary periods. This ancient diversity could still be underestimated due to under-sequencing of the geographic regions concerned by this lineage.

TB-Annotator proved to be useful in the phylogenetic analysis of L5 by providing a finer hierarchical classification and identification of new signature SNPs. The authors note that the phylogenetic resolution is more difficult towards the root, which is expected, but this can be solved with the so-called exclusive variants approach. A notable exception to the overall lack of L5 genomic dynamism is the insertion of IS6110 between spacers 30 and 35 in the CRISPR locus, which is a feature of the L5.1 sublineage.

The study still suffers from some limitations, including a lack of geographic origin information for many samples, absent or scarce data for many countries of Africa, and few genomic studies linking human genetics predisposition to bacterial genetics. However, the authors were able to provide a list of 49 SNPs that allow for the resolution of the phylogenetic tree of L5 with three main branches: L5.1, L5.2, and L5.3 (new). They also linked specific IS6110 insertion/deletion events with particular sublineage emergence, such as the loss of spacers 21-24 in relation to the loss of IS6110 in L5.1.

Finally, the authors discuss the historical timeframe and geographical area of L5 emergence in a wider evolutionary context, suggesting that L5 and L6's highly geographically-constrained patterns may be related to the domestication and migration of *Bos taurus* and *Bos indicus*, respectively. The authors suggest that L4 may have emerged in relation to a North-african *Bos taurus* domestication, and L1-L5-L6 MTBC lineages could have been related to *Bos indicus* domestication and introduction into Africa via East-Africa and the Arabian peninsula. The study provides a more precise picture of some of the evolutionary steps of L5 diversification and offers insight into the intricate relationships between human and animal life-styles, from hunter-gatherer to more recently urban life-styles.

In conclusion, the study offers a comprehensive analysis of L5, providing new insights into its evolutionary history and taxonomy. We were able to use TB-Annotator to provide a finer hierarchical classification and identification of new signature SNPs, which is useful in further studies of tuberculosis caused by *M. africanum*. Although the study still suffers from some limitations, it highlights the need for further genomic studies in underrepresented regions to better understand the virulence evolution of all MTBC lineages.

4 Conclusion

As such, the app is already mature. It incorporates 102,001 strains out of the 167,062 available on NCBI, at the time of this writing. The strains not integrated either have too short reads (the majority) or do not pass our quality filters. In detail, we find a majority of lineage 4 (50%), followed by L2 (33%), L3 and L1 (8% each), the other lineages being marginal (1,122 strains of L6, 452 of L5, 96 of L7, 2 of L8 and 12 of L9). It also contains animal strains, namely *orygis*, *pinipedii*, *microtii*, *dassie* (*Procavia capensis*), *caprae*, *bovis* (including BCG), *mungi* and *chimpanzee*.

The web application contains two main functionalities. The database part allows

multi-criteria searches for the presence of traits of interest: SNP (2,209,467 variants), presence of IS, gene or RD, as well as searches by country, by bioproject, by resistance... The alignment corresponding to the search for these traits can be extracted, and if a tree is built from it by an external application, the newick file can be integrated into the interface in order to exploit the second main functionality, namely visualization and data mining within this tree. By clicking on a strain (SRA), one can visualize all the information about it contained in the database. More interesting, we can make multiple selections (of clades) and see the characters in common and exclusive. We can make multi-criteria searches of characters and visualize in the tree the strains that have them. Finally, we can navigate in the tree (zoom, translation). Because of the various optimization layers mentioned above, all these operations are done in a fluid and almost instantaneous way.

The application is still under intense development. Among the new functionalities being integrated, we can mention the production of an estimate of the spoligotype per SRA (the spacer box is darker the more likely it is to be in the CRISPR locus), the production of a distance matrix per pair of strains in the context of multiple selections, as well as the assembly of reads into contigs. The latter is not currently used, but it should eventually allow the detection of genes that are not present in the H37Rv reference.

For the future, the main objective is to first extend access to the platform to a larger number of private users, and then to move to a web application accessible to all. Such a deployment implies ensuring the scaling up, the security and stability of the platform, as well as its sustainability. Ideally, everyone should be able to deposit their own genomes, with all the guarantees that this implies. Finally, we would like to increase the number of references by not restricting ourselves to H37Rv, integrate long reads (PacBio, etc.), and why not extend the database to other bacteria.

All computations have been performed on the Mésocentre de Franche-Comté supercomputer facilities.

References

1. WHO. The End TB Strategy; 2014. Available from: `https://www.who.int/tb/strategy/en/`.

2. Crubézy E, Ludes B, Poveda JD, Clayton J, Crouau-Roy B, Montagnon D. Identification of Mycobacterium DNA in an Egyptian Pott's disease of 5400 years old. Comptes Rendus de l'Académie des Sciences-Series III-Sciences de la Vie. 1998;321(11):941–951.

3. Cole S, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, et al. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature. 1998;396(6707):190–190.

4. Brites D, Gagneux S. The Nature and Evolution of Genomic Diversity in the Mycobacterium tuberculosis Complex; 2017. p. 1–26. Available from: `http://link.springer.com/10.1007/978-3-319-64371-7_1`.

5. Ngabonziza JCS, Loiseau C, Marceau M, Jouet A, Menardo F, Tzfadia O, et al. A sister lineage of the Mycobacterium tuberculosis complex discovered in the African Great Lakes region. Nature Communications. 2020;11(1):2917. doi:10.1038/s41467-020-16626-6.

6. Coscolla M, Gagneux S, Menardo F, Loiseau C, Ruiz-Rodriguez P, Borrell S, et al. Phylogenomics of Mycobacterium africanum reveals a new lineage and a complex evolutionary history. Microbial Genomics. 2021;7(2). doi:10.1099/mgen.0.000477.

7. Ates LS, Dippenaar A, Ummels R, Piersma SR, van der Woude AD, van Der Kuij K, et al. Mutations in ppe38 block PE_PGRS secretion and increase virulence of Mycobacterium tuberculosis. Nature microbiology. 2018;3(2):181–188.

8. Coll F, Phelan J, Hill-Cawthorne GA, Nair MB, Mallard K, Ali S, et al. Genome-wide analysis of multi-and extensively drug-resistant Mycobacterium tuberculosis. Nature genetics. 2018;50(2):307–316.

9. National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – [cited 2022 Nov 18];. Available from: `https://www.ncbi.nlm.nih.gov/`.

10. Phelan JE, O'Sullivan DM, Machado D, Ramos J, Oppong YE, Campino S, et al. Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. Genome medicine. 2019;11(1):1–7.

11. Feuerriegel S, Schleusener V, Beckert P, Kohl TA, Miotto P, Cirillo DM, et al. PhyResSE: a Web tool delineating Mycobacterium tuberculosis antibiotic resistance and lineage from whole-genome sequencing data. Journal of clinical microbiology. 2015;53(6):1908–1914.

12. Guyeux C, Senelle G, Refrégier G, Cambau E, Sola C. Description of a new ancestral lineage of the L2 Mycobacterium tuberculosis complex in Japan and Revision of the current L2 Lineage/sublineage nomenclature; 2021.

13. Senelle G, Sahal MR, La K, Molina-Moya B, Dominguez J, Panda T, et al. An updated evolutionary history and taxonomy of Mycobacterium tuberculosis lineage 5, also called M. africanum. bioRxiv. 2022;doi:10.1101/2022.11.21.517336.

14. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018;34(23):4121–4123.

15. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, et al. Sustainable data analysis with Snakemake. F1000Research. 2021;10:33. doi:10.12688/f1000research.29032.1.

16. Kubernetes. Kubernetes;. Available from: `https://kubernetes.io/fr/`.

17. Argo. Argo Workflows;. Available from: `https://github.com/argoproj/argo-workflows`.

18. Leinonen R, Sugawara H, Shumway M. The Sequence Read Archive. Nucleic Acids Research. 2011;39:D19–D21. doi:10.1093/nar/gkq1019.

19. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research. 2010;38(6):1767–1771. doi:10.1093/nar/gkp1137.

20. Shen W, Le S, Li Y, Hu F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLOS ONE. 2016;11:e0163962. doi:10.1371/journal.pone.0163962.

21. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884–i890. doi:10.1093/bioinformatics/bty560.

22. Kivioja T, Vähärautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, et al. Counting absolute numbers of molecules using unique molecular identifiers. Nature methods. 2012;9(1):72–74.

23. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM; 2013.

24. Liu Y, Popp B, Schmidt B. CUSHAW3: Sensitive and Accurate Base-Space and Color-Space Short-Read Alignment with Hybrid Seeding. PLoS ONE. 2014;9(1):e86869. doi:10.1371/journal.pone.0086869.

25. Highnam G, Wang JJ, Kusler D, Zook J, Vijayan V, Leibovich N, et al. An analytical framework for optimizing variant discovery from personal genomes. Nature Communications. 2015;6(1):6275. doi:10.1038/ncomms7275.

26. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–2079. doi:10.1093/bioinformatics/btp352.

27. Ebbert MTW, Wadsworth ME, Staley LA, Hoyt KL, Pickett B, Miller J, et al. Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. BMC Bioinformatics. 2016;17(S7):239. doi:10.1186/s12859-016-1097-3.

28. Hsi-Yang Fritz M, Leinonen R, Cochrane G, Birney E. Efficient storage of high throughput DNA sequencing data using reference-based compression. Genome Research. 2011;21(5):734–740. doi:10.1101/gr.114819.110.

29. Seemann T. snippy: fast bacterial variant calling from NGS reads; 2015. Available from: `https://github.com/tseemann/snippy`.

30. Bush SJ, Foster D, Eyre DW, Clark EL, De Maio N, Shaw LP, et al. Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism–calling pipelines. GigaScience. 2020;9(2). doi:10.1093/gigascience/giaa007.

31. Seemann T. Samclip: filter SAM file for soft and hard clipped alignments; 2020. Available from: `https://github.com/tseemann/samclip`.

32. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing; 2012.

33. Meehan CJ, Goig GA, Kohl TA, Verboven L, Dippenaar A, Ezewudo M, et al. Whole genome sequencing of Mycobacterium tuberculosis: current standards and open issues. Nature Reviews Microbiology. 2019;17(9):533–545. doi:10.1038/s41579-019-0214-5.

34. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly;6:80–92. doi:10.4161/fly.19695.

35. Holmes JB, Moyer E, Phan L, Maglott D, Kattman B. SPDI: data model for variants and applications at NCBI. Bioinformatics. 2020;36:1902–1907. doi:10.1093/bioinformatics/btz856.

36. Suzuki S, Yasuda T, Shiraishi Y, Miyano S, Nagasaki M. ClipCrop: a tool for detecting structural variations with single-base resolution using soft-clipping information. BMC Bioinformatics. 2011;12(Suppl 14):S7. doi:10.1186/1471-2105-12-S14-S7.

37. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–842. doi:10.1093/bioinformatics/btq033.

38. Constant P, Perez E, Malaga W, Lanéelle MA, Saurel O, Daffé M, et al. Role of the pks15/1 Gene in the Biosynthesis of Phenolglycolipids in the Mycobacterium tuberculosisComplex. Journal of Biological Chemistry. 2002;277(41):38148–38158.

39. Faksri K, Xia E, Tan JH, Teo YY, Ong RTH. In silico region of difference (RD) analysis of Mycobacterium tuberculosis complex from sequence reads using RD-Analyzer. BMC Genomics. 2016;17(1):847. doi:10.1186/s12864-016-3213-1.

40. Zimpel CK, Patané JSL, Guedes ACP, de Souza RF, Silva-Pereira TT, Camargo NCS, et al. Global Distribution and Evolution of Mycobacterium bovis Lineages. Frontiers in Microbiology. 2020;11. doi:10.3389/fmicb.2020.00843.

41. Reddy TBK, Riley R, Wymore F, Montgomery P, DeCaprio D, Engels R, et al. TB database: an integrated platform for tuberculosis research. Nucleic Acids Research. 2009;37:D499–D508. doi:10.1093/nar/gkn652.

42. WHO. Catalogue of mutations in Mycobacterium tuberculosis complex and their association with drug resistance; 2021. Available from: `https://www.who.int/publications/i/item/978924002817`.

43. Coll F, McNerney R, Guerra-Assunção JA, Glynn JR, Perdigão J, Viveiros M, et al. A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. Nature communications. 2014;5(1):1–5.

44. Palittapongarnpim P, Ajawatanawong P, Viratyosin W, Smittipat N, Disratthakit A, Mahasirimongkol S, et al. Evidence for host-bacterial co-evolution via genome sequence analysis of 480 Thai Mycobacterium tuberculosis lineage 1 isolates. Scientific reports. 2018;8(1):11597.

45. Lipworth S, Jajou R, De Neeling A, Bradley P, Van Der Hoek W, Maphalala G, et al. SNP-IT tool for identifying subspecies and associated lineages of Mycobacterium tuberculosis complex. Emerging Infectious Diseases. 2019;25(3):482.

46. Napier G, Campino S, Merid Y, Abebe M, Woldeamanuel Y, Aseffa A, et al. Robust barcoding and identification of Mycobacterium tuberculosis lineages for epidemiological and clinical studies. Genome medicine. 2020;12:1–10.

47. Freschi L, Vargas Jr R, Husain A, Kamal SM, Skrahina A, Tahseen S, et al. Population structure, biogeography and transmissibility of Mycobacterium tuberculosis. Nature communications. 2021;12(1):6099.

48. Coscolla M, Gagneux S, Menardo F, Loiseau C, Ruiz-Rodriguez P, Borrell S, et al. Phylogenomics of Mycobacterium africanum reveals a new lineage and a complex evolutionary history. Microbial genomics. 2021;7(2).

49. Coll F, McNerney R, Preston MD, Guerra-Assunção JA, Warry A, Hill-Cawthorne G, et al. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. Genome medicine. 2015;7(1):1–11.

50. Yoshida SY, Suzuki Y, Kishi F, Wada T, Hasegawa N, Ide K, 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. 2017;107:125–133.

51. Bradley P, Gordon NC, Walker TM, Dunn L, Heys S, Huang B, et al. Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. Nature Communications. 2015;6(1):1–10.

52. Allix-Béguec C, Harmsen D, Weniger T, Supply P, Niemann S. Evaluation and user-strategy of MIRU-VNTRplus, a multifunctional database for online analysis of genotyping data and phylogenetic identification of Mycobacterium tuberculosis complex isolates. Journal of Clinical Microbiology. 2008;46(8):2692–2699.

53. Demay C, Liens B, Burguière T, Hill V, Couvin D, Millet J, et al. SITVITWEB–a publicly available international multimarker database for studying Mycobacterium tuberculosis genetic diversity and molecular epidemiology. Infection, Genetics and Evolution. 2012;12(4):755–766.

54. Consortium T. TBDB: Tuberculosis Database; 2013. `http://www.tbdb.org/`.

55. Consortium T. TubercuList: Mycobacterium tuberculosis H37Rv Database; 2010. `http://tuberculist.epfl.ch/`.

56. Kapopoulou A, Lew JM, Cole ST. Mycobrowser: A web-based resource for tuberculosis research. Nucleic acids research. 2011;39(suppl_1):D633–D636.

57. Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T, Bun C, et al. Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. Nucleic acids research. 2017;45(D1):D535–D542.

58. Karp PD, Weaver D, Latendresse M, Dräger A, Paley S. BioCyc: a reference collection of pathway/genome databases. Nucleic acids research. 2021;49(D1):D743–D750.

59. Consortium E. Enterobase: Bacterial genomic data and analysis; 2017. `https://enterobase.warwick.ac.uk/species/index/mycobacterium`.

60. Consortium P. PathogenSeq: Tools for bacterial genomic analysis; 2017. `http://pathogenseq.lshtm.ac.uk/#portfolioTB6`.

61. Consortium C. CNGb Pathogen Variation Database; 2021. `https://db.cngb.org/pvd/global-result/?global=tuberculosis`.

62. Guyeux C, Senelle G, Refrégier G, Bretelle-Establet F, Cambau E, Sola C. Connection between two historical tuberculosis outbreak sites in Japan, Honshu, by a new ancestral Mycobacterium tuberculosis L2 sublineage. Epidemiology & Infection. 2022;150.

Annexe B

Tableaux

Référence	Début (incl.)	Fin (excl.)	Nom	Notes
NC_000962.3	4350265	4359722	RD1	Mahairas et al. [172]
NC_000962.3	2221059	2231846	RD2	Mahairas et al. [172]
NC_000962.3	1779266	1788513	RD3	Mahairas et al. [172]
NC_000962.3	1696016	1708748	RD4	Brosch et al. [25]
NC_000962.3	2626069	2635032	RD5	Gordon et al. [100]
NC_000962.3	2208005	2220724	RD7	Gordon et al. [100]
NC_000962.3	4056840	4062734	RD8	Gordon et al. [100]
NC_000962.3	2330073	2332102	RD9	Gordon et al. [100]
NC_000962.3	264754	266656	RD10	Gordon et al. [100]
NC_000962.3	2970016	2980998	RD11	Behr [16]
NC_000962.3	3479430	3491865	RD12can	Marmiesse et al. [174]
NC_000962.3	3484740	3487515	RD12	Behr [16]
NC_000962.3	1402930	1405937	RD13	Behr [16]
NC_000962.3	1998225	2007293	RD14	Behr [16]
NC_000962.3	2220947	2222888	RD2seal	Marmiesse et al. [174]
NC_000962.3	79569	83035	105	Tsolaki et al. [281]
NC_000962.3	96642	99258	108	Tsolaki et al. [281]
NC_000962.3	229870	231700	110	Tsolaki et al. [281]
NC_000962.3	466047	468299	116	Tsolaki et al. [281]
NC_000962.3	483943	490774	117	Tsolaki et al. [281]
NC_000962.3	608541	610959	120	Tsolaki et al. [281]
NC_000962.3	663096	665816	121	Tsolaki et al. [281]
NC_000962.3	669795	670965	122	Tsolaki et al. [281]
NC_000962.3	746520	747806	127	Tsolaki et al. [281]
NC_000962.3	769382	770003	129	Tsolaki et al. [281]
NC_000962.3	840970	841928	130	Tsolaki et al. [281]
NC_000962.3	944126	949488	132	Tsolaki et al. [281]
NC_000962.3	1074125	1075947	134	Tsolaki et al. [281]
NC_000962.3	1332184	1335034	142	Tsolaki et al. [281]
NC_000962.3	1519819	1524571	145	Tsolaki et al. [281]
NC_000962.3	1896864	1899350	150	Tsolaki et al. [281]
NC_000962.3	2127983	2128973	163	Tsolaki et al. [281]
NC_000962.3	2139723	2140030	164	Tsolaki et al. [281]
NC_000962.3	2179564	2180649	167	Tsolaki et al. [281]
NC_000962.3	2198474	2200489	168	Tsolaki et al. [281]
NC_000962.3	2361909	2363682	178	Tsolaki et al. [281]
NC_000962.3	2535431	2536141	181	Tsolaki et al. [281]

Référence	Début (incl.)	Fin (excl.)	Nom	Notes
NC_000962.3	3106811	3107655	202	Tsolaki et al. [281]
NC_000962.3	3111820	3115333	203	Tsolaki et al. [281]
NC_000962.3	3119370	3119759	206	Tsolaki et al. [281]
NC_000962.3	3448507	3451398	219	Tsolaki et al. [281]
NC_000962.3	3551230	3555383	223	Tsolaki et al. [281]
NC_000962.3	3868780	3869670	231	Tsolaki et al. [281]
NC_000962.3	4092080	4092921	239	Tsolaki et al. [281]
NC_000962.3	4209438	4211501	246	Tsolaki et al. [281]
NC_000962.3	254002	257597	110a	Tsolaki et al. [281]
NC_000962.3	289423	289885	110b	Tsolaki et al. [281]
NC_000962.3	289977	290080	110c	Tsolaki et al. [281]
NC_000962.3	888347	888787	131ab	Tsolaki et al. [281]
NC_000962.3	889017	890375	131e	Tsolaki et al. [281]
NC_000962.3	1243196	1243873	141a	Tsolaki et al. [281]
NC_000962.3	1526657	1529303	145a	Tsolaki et al. [281]
NC_000962.3	1691100	1694978	147b	Tsolaki et al. [281]
NC_000962.3	2153061	2153385	165	Tsolaki et al. [281]
NC_000962.3	2153076	2161126	166	Tsolaki et al. [281]
NC_000962.3	2235690	2238063	172	Tsolaki et al. [281]
NC_000962.3	2235722	2237377	172a	Tsolaki et al. [281]
NC_000962.3	2237051	2240700	174	Tsolaki et al. [281]
NC_000962.3	2245205	2248533	174a	Tsolaki et al. [281]
NC_000962.3	2263450	2263638	175a	Tsolaki et al. [281]
NC_000962.3	2545196	2551675	182	Tsolaki et al. [281]
NC_000962.3	2546465	2548425	182a	Tsolaki et al. [281]
NC_000962.3	2866469	2867124	196	Tsolaki et al. [281]
NC_000962.3	2867121	2867777	196b	Tsolaki et al. [281]
NC_000962.3	3120524	3127922	207	Tsolaki et al. [281]
NC_000962.3	3120431	3123025	210	Tsolaki et al. [281]
NC_000962.3	3288359	3290567	213a	Tsolaki et al. [281]
NC_000962.3	4212651	4215045	247	Tsolaki et al. [281]
NC_000962.3	4212914	4215052	247b	Tsolaki et al. [281]
NC_000962.3	4370422	4373232	252	Tsolaki et al. [281]
NC_000962.3	4371818	4373776	252b	Tsolaki et al. [281]
NC_000962.3	4056662	4058980	236	Tsolaki et al. [281]
NC_000962.3	4056945	4058395	236a	Tsolaki et al. [281]
NC_000962.3	2969989	2980970	198a	Tsolaki et al. [281]
NC_000962.3	2704308	2704808	193	Tsolaki et al. [281]
NC_000962.3	2631497	2636953	188	Tsolaki et al. [281]
NC_000962.3	2585855	2588771	183	Tsolaki et al. [281]
NC_000962.3	2225940	2228588	171	Tsolaki et al. [281]
NC_000962.3	1986638	1998622	152	Tsolaki et al. [281]
NC_000962.3	1779266	1788513	149	Tsolaki et al. [281]
NC_000962.3	1718910	1721212	147c	Tsolaki et al. [281]
NC_000962.3	453366	455972	115	Tsolaki et al. [281]
NC_000962.3	170008	173787	109c	Tsolaki et al. [281]
NC_000962.3	170014	173791	DS1	Kato-Maeda et al. [144]
NC_000962.3	453366	455972	DS2	Kato-Maeda et al. [144]
NC_000962.3	886542	887416	DS41	Kato-Maeda et al. [144]
NC_000962.3	931783	932201	DS47	Kato-Maeda et al. [144]
NC_000962.3	1310758	1311678	DS16	Kato-Maeda et al. [144]

Référence	Début (incl.)	Fin (excl.)	Nom	Notes
NC_000962.3	1718912	1721221	DS4	Kato-Maeda et al. [144]
NC_000962.3	1727660	1728463	DS32	Kato-Maeda et al. [144]
NC_000962.3	1779278	1788513	DS5	Kato-Maeda et al. [144]
NC_000962.3	1986625	1987702	DS6L	Kato-Maeda et al. [144]
NC_000962.3	1543306	1998604	DS6	Kato-Maeda et al. [144]
NC_000962.3	2225939	2228589	DS7	Kato-Maeda et al. [144]
NC_000962.3	2381414	2383685	DS20	Kato-Maeda et al. [144]
NC_000962.3	2585855	2588771	DS8	Kato-Maeda et al. [144]
NC_000962.3	2627268	2632931	DS9	Kato-Maeda et al. [144]
NC_000962.3	2704309	2704807	DS33	Kato-Maeda et al. [144]
NC_000962.3	2969984	2980971	DS10	Kato-Maeda et al. [144]
NC_000962.3	3120467	3123066	DS71	Kato-Maeda et al. [144]
NC_000962.3	3448496	3451387	DS19	Kato-Maeda et al. [144]
NC_000962.3	3842309	3847235	DS13	Kato-Maeda et al. [144]
NC_000962.3	3868780	3869670	DS26	Kato-Maeda et al. [144]
NC_000962.3	3955467	3956104	DS27	Kato-Maeda et al. [144]
NC_000962.3	4056945	4058396	DS21	Kato-Maeda et al. [144]
NC_000962.3	4212650	4215046	DS18	Kato-Maeda et al. [144]
NC_000962.3	4370423	4373246	DS15	Kato-Maeda et al. [144]
NC_000962.3	4386640	4388340	DS70	Kato-Maeda et al. [144]
NC_000962.3	1153022	1156703	139BWa	Tsolaki et al. [282]
NC_000962.3	1498104	1499473	144BWa	Tsolaki et al. [282]
NC_000962.3	2190643	2194795	168BWa	Tsolaki et al. [282]
NC_000962.3	3004384	3006633	200BWa	Tsolaki et al. [282]
NC_000962.3	3292402	3294519	213BWa	Tsolaki et al. [282]
NC_000962.3	3489098	3489320	220BWa	Tsolaki et al. [282]
NC_000962.3	3683271	3686635	224ca	Tsolaki et al. [282]
NC_000962.3	4340417	4354536	RD1mic	Brodin et al. [23]
NC_000962.3	2627831	2635576	RD5mic	Brodin et al. [23]
NC_000962.3	3121879	3126682	MiD1	Brodin et al. [23]
NC_000962.3	3554067	3555259	MiD2	Brodin et al. [23]
NC_000962.3	3741140	3755776	MiD3	Brodin et al. [23]
NC_000962.3	3477171	3492150	RD12riyadh	Guan et al. [103]
NC_000962.3	1332921	1334468	N-RD18	Salamon [232]
NC_000962.3	3897072	3897787	N-RD17	Salamon [232]
NC_000962.3	4189607	4190760	N-RD25	Salamon [232]
NC_000962.3	76526	84934	RD720	Mostowy et al. [192]
NC_000962.3	82922	84589	RD721	Mostowy et al. [192]
NC_000962.3	149240	150980	RD701	Mostowy et al. [192]
NC_000962.3	216794	218516	RD702	Mostowy et al. [192]
NC_000962.3	1041192	1049812	RD722	Mostowy et al. [192]
NC_000962.3	1501712	1503655	RD711	Mostowy et al. [192]
NC_000962.3	1875725	1881660	RD742	Mostowy et al. [192]
NC_000962.3	2219418	2223186	RD713	Mostowy et al. [192]
NC_000962.3	2235803	2239118	RD743	Mostowy et al. [192]
NC_000962.3	2265111	2266239	RD724	Mostowy et al. [192]
NC_000962.3	2629496	2634359	RD728	Mostowy et al. [192]
NC_000962.3	2784616	2785969	RD715	Mostowy et al. [192]
NC_000962.3	2902566	2904318	RD727	Mostowy et al. [192]
NC_000962.3	3781987	3782408	RD735	Mostowy et al. [192]
NC_000962.3	3904957	3906706	726	Gagneux et al. [91]

Référence	Début (incl.)	Fin (excl.)	Nom	Notes
NC_000962.3	1710766	1711556	750	Gagneux et al. [91]
NC_000962.3	1502786	1503881	761	Gagneux et al. [91]
NC_000962.3	29987	34321	RDcap_Spain1	Mostowy et al. [193]
NC_000962.3	136717	138162	RDcap_Spain2	Mostowy et al. [193]
NC_000962.3	188884	190192	RDcap_Spain3	Mostowy et al. [193]
NC_000962.3	484395	485447	RDbovis_buff1	Mostowy et al. [193]
NC_000962.3	533231	534212	RDbovis_sigK	Mostowy et al. [193]
NC_000962.3	669912	683321	RDbovis_1160_1	Mostowy et al. [193]
NC_000962.3	1151878	1154484	RDbovis_kdp	Mostowy et al. [193]
NC_000962.3	1376214	1376863	RDcap_Asia1	Mostowy et al. [193]
NC_000962.3	1523188	1547065	RDbovis_Kruger	Mostowy et al. [193]
NC_000962.3	1720073	1720718	RDbovis_wbbl2	Mostowy et al. [193]
NC_000962.3	1987197	1998792	RDoryx_wag22*	Mostowy et al. [193]
NC_000962.3	1996617	1999707	RDcap_Asia2	Mostowy et al. [193]
NC_000962.3	2038716	2048278	RDoryx_1*	Mostowy et al. [193]
NC_000962.3	2180593	2187400	RDbovis_0173	Mostowy et al. [193]
NC_000962.3	2195376	2198580	RDcap_Spain4	Mostowy et al. [193]
NC_000962.3	2359036	2362916	RDcap_Asia3	Mostowy et al. [193]
NC_000962.3	2629040	2639540	RD5oryx*	Mostowy et al. [193]
NC_000962.3	3119191	3120523	RDcap_Spain5	Mostowy et al. [193]
NC_000962.3	3447447	3451241	RDbovis_virS	Mostowy et al. [193]
NC_000962.3	3477735	3482410	RD12HUP	Mostowy et al. [193]
NC_000962.3	3479669	3491251	RD12oryx*	Mostowy et al. [193]
NC_000962.3	3549074	3555365	RDoryx_4*	Mostowy et al. [193]
NC_000962.3	3823566	3825360	RDbovis_1160_2	Mostowy et al. [193]
NC_000962.3	3945596	3951328	RDbovis_fadD18	Mostowy et al. [193]
NC_000962.3	3967679	3968974	RDpin	Mostowy et al. [193]
NC_000962.3	4142143	4143783	RDbovis_buff2	Mostowy et al. [193]
NC_000962.3	4194727	4196290	RDcap_Spain6	Mostowy et al. [193]
NC_000962.3	4366791	4376361	RDcap_Spain7	Mostowy et al. [193]
NC_000962.3	4370864	4375132	RDbovis_Δpan	Mostowy et al. [193]
NC_000962.3	664279	669600	RDAf1	Müller et al. [196]
NC_000962.3	680336	694428	RDAf2	Berg et al. [17]
NC_000962.3	1768073	1768877	RDEu1	Smith et al. [250]
NC_000962.3	4371019	4373425	RDpan	Rauzier et al. [221]
NC_000962.3	2627068	2636921	RD5das	Mostowy et al. [191]
NC_000962.3	3447250	3448436	RDVirsdas	Mostowy et al. [191]
NC_000962.3	4188300	4190242	N-RD25das	Mostowy et al. [191]
NC_000962.3	4352302	4356435	RD1das	Mostowy et al. [191]
NC_000962.3	3296370	3296371	pks 1/15 7bp	Constant et al. [56]
NC_000962.3	3296370	3296371	pks 1/15 6bp	Marmiesse et al. [174]
NC_000962.3	2784160	2784568	CUS_GS_RD715	
NC_000962.3	2949905	2955132	RD197	Schürch et al. [233]
NC_000962.3	2626968	2633061	RD185	Schürch et al. [233]
NC_000962.3	358029	363748	RD112	Schürch et al. [233]
NC_000962.3	1715869	1733378	RD148	Schürch et al. [233]
NC_000962.3	2238646	2242137	RD174	Schürch et al. [233]
NC_000962.3	2128378	2129584	RD163	Schürch et al. [233]
NC_000962.3	2038670	2055881	RDsurRv1799	Dippenaar et al. [75]
NC_000962.3	2163477	2166383	RDsurppe34	Dippenaar et al. [75]
NC_000962.3	2534706	2534981	RDsurRv2262	Dippenaar et al. [75]

Référence	Début (incl.)	Fin (excl.)	Nom	Notes
NC_000962.3	2578443	2578536	RDsurRv2307c	Dippenaar et al. [75]
NC_000962.3	2627066	2636919	RD5sur	Dippenaar et al. [75]
NC_000962.3	3113849	3127190	RDsurDR	Dippenaar et al. [75]
NC_000962.3	3710233	3710381	RDsurRv3324A	Dippenaar et al. [75]
NC_000962.3	2214297	2217261	RD7tb	Huard et al. [125]
NC_000962.3	4056945	4058397	RD236a	Huard et al. [125]
NC_000962.3	4189257	4190367	N-RD25tbA	Huard et al. [125]
NC_000962.3	4189233	4190191	N-RD25tbB	Huard et al. [125]
NC_000962.3	4368661	4368718	RD1tbB	Huard et al. [125]
NC_000962.3	4352092	4353695	RD1mon	Alexander et al. [5]
NC_000962.3	34788	35208	RD301	Bespiatykh et al. [18]
NC_000962.3	164949	165738	RD302	Bespiatykh et al. [18]
NC_000962.3	321665	322040	RD303	Bespiatykh et al. [18]
NC_000962.3	426726	428184	RD304	Bespiatykh et al. [18]
NC_000962.3	472035	472534	RD305	Bespiatykh et al. [18]
NC_000962.3	1313127	1313384	RD317	Bespiatykh et al. [18]
NC_000962.3	1313134	1313390	RD306	Bespiatykh et al. [18]
NC_000962.3	1714722	1719659	RD307	Bespiatykh et al. [18]
NC_000962.3	1727147	1728126	RD308	Bespiatykh et al. [18]
NC_000962.3	1897541	1901104	RD309	Bespiatykh et al. [18]
NC_000962.3	2197955	2198334	RD310	Bespiatykh et al. [18]
NC_000962.3	2729620	2729833	RD311	Bespiatykh et al. [18]
NC_000962.3	2784161	2784568	RD312	Bespiatykh et al. [18]
NC_000962.3	3037025	3037297	RD313	Bespiatykh et al. [18]
NC_000962.3	3606557	3607839	RD314	Bespiatykh et al. [18]
NC_000962.3	3868985	3869341	RD315	Bespiatykh et al. [18]
NC_000962.3	3952928	3954225	RD316	Bespiatykh et al. [18]
NC_000962.3	76163	84825	RD105ext	Shitikov et al. [244]
NC_000962.3	3743197	3769514	RDRio	Lazzarini et al. [156]
NC_000962.3	3378269	3380708	MiD4	Garcia-Pelayo et al. [95]
NC_000962.3	2562514	2563875	ND1	
NC_000962.3	3883907	3887201	ND2	

TABLE B.1 – Tableau récapitulatif des RD mis à jour sur la référence NC_000962.3. Extraction des positions effectuée via jonction PCR et mapping. Les positions sont données en base-0.

Nom du champ
geo_loc_name
country
Country
Isolation_country
geo loc name
geographic location (country and/or sea region)
geographic location (country and/or sea region)
geographic location (country and/or sea)
geographic location (country and/or sea region)
geographic location (country and/or searegion)
geographic location (country)
geographic location (country :regionarea)
geographic location (locality)
geographic location (region and locality)
geographic location country and or sea
Geographic location name country
geographic locations
geographic location
geographic origin
geographical location
geographical location (country :region location)
geolocname
lat_lon
Lat_lon
latitude and longitude
geographic location (latitude and longitude)
geographic location (latitude longitude)
geographical location (lat lon)
geographical location (longitude and longitude)
lat lon
latlon

TABLE B.2 – Liste des champs de métadonnées de géolocalisation du NCBI normalisés par TB-Annotator

Métadonnée
Nom scientifique
Pays
Coordonnées géographiques
Hôte
Source d'isolation
Date de collecte
Status VIH
Status culture

TABLE B.3 – Liste des métadonnées extraites normalisées supportées par TB-Annotator

Valeur
Homo sapiens
Bovinae
Suidae
Meles
Ferret
Possum
Caprinae
Raccoon
Muridae
Elephantidae
Felinae
Dassie
Primate
Antilopinae
Sea lion
Canidae

Table B.4 – Liste des valeurs de métadonnées d'hôte supportées par TB-Annotator

Valeur
Sputum
Tissue
Pulmonary
Simulated
Lymph node
Cerebrospinal fluid
Pleural fluid

Table B.5 – Liste des valeurs de métadonnées de source d'isolation supportées par TB-Annotator

Table des figures

Liste des tableaux

Bibliographie

[1] Efficacy, safety and immunogenicity evaluation of mtbvac in newborns in sub-saharan africa (mtbvacn3). URL https://www.clinicaltrials.gov/study/NCT04975178.

[2] National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD) : National library of medicine (us), national center for biotechnology information ; [1988] – [cited 2022 nov 18]. Available from : https://www.ncbi.nlm.nih.gov/.

[3] Dissou Affolabi, Frank Faïhun, N'dira Sanoussi, Gladys Anyo, Isdore Chola Shamputa, Leen Rigouts, Luc Kestens, Séverin Anagonou, and Françoise Portaels. Possible outbreak of streptomycin-resistant mycobacterium tuberculosis beijing in benin. *Emerg. Infect. Dis.*, 15(7) :1123–1125, July 2009.

[4] Pravech Ajawatanawong, Hideki Yanai, Nat Smittipat, Areeya Disratthakit, Norio Yamada, Reiko Miyahara, Supalert Nedsuwan, Worarat Imasanguan, Pacharee Kantipong, Boonchai Chaiyasirinroje, Jiraporn Wongyai, Supada Plitphonganphim, Pornpen Tantivitayakul, Jody Phelan, Julian Parkhill, Taane G Clark, Martin L Hibberd, Wuthiwat Ruangchai, Panawun Palittapongarnpim, Tada Juthayothin, Yuttapong Thawornwattana, Wasna Viratyosin, Sissades Tongsima, Surakameth Mahasirimongkol, Katsushi Tokunaga, and Prasit Palittapongarnpim. A novel ancestral beijing sublineage of mycobacterium tuberculosis suggests the transition site to modern beijing sublineages. *Sci. Rep.*, 9(1) :13718, September 2019.

[5] Kathleen A. Alexander, Claire E. Sanderson, Michelle H. Larsen, Suelee Robbe-Austerman, Mark C. Williams, and Mitchell V. Palmer. Emerging tuberculosis pathogen hijacks social communication behavior in the group-living banded mongoose (mungos mungo). *mBio*, 7, 7 2016. ISSN 2161-2129. doi : 10.1128/mBio.00281-16.

[6] Caroline Allix-Béguec, Dag Harmsen, Thomas Weniger, Philip Supply, and Stefan Niemann. 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(8) :2692–2699, August 2008.

[7] Caroline Allix-Béguec, Dag Harmsen, Thomas Weniger, Philip Supply, and Stefan Niemann. Evaluation and user-strategy of miru-vntrplus, a multifunctional database for online analysis of genotyping data and phylogenetic identification of mycobacterium tuberculosis complex isolates. *Journal of Clinical Microbiology*, 46(8) : 2692–2699, 2008.

[8] Caroline Allix-Béguec, Céline Wahl, Madeleine Hanekom, Vladyslav Nikolayevskyy, Francis Drobniewski, Shinji Maeda, Isolina Campos-Herrero, Igor Mokrousov, Stefan Niemann, Irina Kontsevaya, Nalin Rastogi, Sofia Samper, Li-Hwei Sng, Robin M Warren, and Philip Supply. 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(1) :164–172, January 2014.

[9] Gil Amitai and Rotem Sorek. Crispr–cas adaptation : insights into the mechanism of action. *Nature Reviews Microbiology*, 14 :67–76, 2 2016. ISSN 1740-1526. doi : 10.1038/nrmicro.2015.14.

[10] Argo. Argo Workflows. URL https://github.com/argoproj/argo-workflows.

[11] Oussama Baker, Oona Y.-C. Lee, Houdini H.T. Wu, Gurdyal S. Besra, David E. Minnikin, Gareth Llewellyn, Christopher M. Williams, Frank Maixner, Niall O'Sullivan, Albert Zink, Bérénice Chamel, Rima Khawam, Eric Coqueugniot, Daniel Helmer, Françoise Le Mort, Pascale Perrin, Lionel Gourichon, Bruno Dutailly, György Pálfi, Hélène Coqueugniot, and Olivier Dutour. Human tuberculosis predates domestication in ancient syria. *Tuberculosis*, 95 :S4–S12, 6 2015. ISSN 14729792. doi : 10.1016/j.tube.2015.02.001.

[12] Francois Balloux and Lucy van Dorp. Q&a : What are pathogens, and what have they done to and for us ? *BMC Biology*, 15 :91, 12 2017. ISSN 1741-7007. doi : 10.1186/s12915-017-0433-z.

[13] Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, Sergey I. Nikolenko, Son Pham, Andrey D. Prjibelski, Alexey V. Pyshkin, Alexander V. Sirotkin, Nikolay Vyahhi, Glenn Tesler, Max A. Alekseyev, and Pavel A. Pevzner. Spades : A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, 19 :455–477, 5 2012. ISSN 1066-5277. doi : 10.1089/cmb.2012.0021.

[14] I Barberis, N L Bragazzi, L Galluzzo, and M Martini. The history of tuberculosis : from the first historical records to the isolation of koch's bacillus. *Journal of preventive medicine and hygiene*, 58 :E9–E12, 3 2017. ISSN 1121-2233.

[15] Rodolphe Barrangou, Christophe Fremaux, Hélène Deveau, Melissa Richards, Patrick Boyaval, Sylvain Moineau, Dennis A. Romero, and Philippe Horvath. Crispr provides acquired resistance against viruses in prokaryotes. *Science*, 315 :1709–1712, 3 2007. ISSN 0036-8075. doi : 10.1126/science.1138140.

[16] M. A. Behr. Comparative Genomics of BCG Vaccines by Whole-Genome DNA Microarray. *Science*, 284(5419) :1520–1523, may 1999. ISSN 00368075. doi : 10.1126/science.284.5419.1520. URL https://www.sciencemag.org/lookup/doi/10.1126/science.284.5419.1520.

[17] Stefan Berg, M. Carmen Garcia-Pelayo, Borna Müller, Elena Hailu, Benon Asiimwe, Kristin Kremer, James Dale, M. Beatrice Boniotti, Sabrina Rodriguez, Markus Hilty, Leen Rigouts, Rebuma Firdessa, Adelina Machado, Custodia Mucavele, Bongo Nare Richard Ngandolo, Judith Bruchfeld, Laura Boschiroli, Annélle Müller, Naima Sahraoui, Maria Pacciarini, Simeon Cadmus, Moses Joloba, Dick van Soolingen, Anita L. Michel, Berit Djønne, Alicia Aranaz, Jakob Zinsstag, Paul van Helden, Françoise Portaels, Rudovick Kazwala, Gunilla Källenius, R. Glyn Hewinson, Abraham Aseffa, Stephen V. Gordon, and Noel H. Smith. African 2, a clonal complex of mycobacterium bovis epidemiologically important in east africa. *Journal of Bacteriology*, 193 :670–678, 2 2011. ISSN 0021-9193. doi : 10.1128/JB.00750-10.

[18] D Bespiatykh, J Bespyatykh, I Mokrousov, and E Shitikov. A comprehensive map of mycobacterium tuberculosis complex regions of difference. *mSphere*, 6 :e0053521, 8 2021. ISSN 2379-5042. doi : 10.1128/mSphere.00535-21.

[19] Pablo J Bifani, Barun Mathema, Natalia E Kurepina, and Barry N Kreiswirth. Global dissemination of the mycobacterium tuberculosis W-Beijing family strains. *Trends Microbiol.*, 10(1) :45–52, January 2002.

[20] Kirsten I. Bos, Kelly M. Harkins, Alexander Herbig, Mireia Coscolla, Nico Weber, Iñaki Comas, Stephen A. Forrest, Josephine M. Bryant, Simon R. Harris, Verena J. Schuenemann, Tessa J. Campbell, Kerttu Majander, Alicia K. Wilbur, Ricardo A. Guichon, Dawnie L. Wolfe Steadman, Della Collins Cook, Stefan Niemann, Marcel A. Behr, Martin Zumarraga, Ricardo Bastida, Daniel Huson, Kay Nieselt, Douglas Young, Julian Parkhill, Jane E. Buikstra, Sebastien Gagneux, Anne C. Stone, and Johannes Krause. Pre-columbian mycobacterial genomes reveal seals as a source of new world human tuberculosis. *Nature*, 514 :494–497, 10 2014. ISSN 0028-0836. doi : 10.1038/nature13591.

[21] Daria Bottai, Wafa Frigui, Fadel Sayes, Mariagrazia Di Luca, Dalila Spadoni, Alexandre Pawlik, Marina Zoppo, Mickael Orgeur, Varun Khanna, David Hardy, Sophie Mangenot, Valerie Barbe, Claudine Medigue, Laurence Ma, Christiane Bouchier, Arianna Tavanti, Gerald Larrouy-Maumus, and Roland Brosch. TbD1 deletion as a driver of the evolutionary success of modern epidemic Mycobacterium tuberculosis lineages. *Nature Communications*, 11(1) :684, dec 2020. ISSN 2041-1723. doi : 10.1038/s41467-020-14508-5. URL http://www.nature.com/articles/s41467-020-14508-5.

[22] Phelim Bradley, N Claire Gordon, Timothy M Walker, Laura Dunn, Simon Heys, Becca Huang, Sarah Earle, Louise J Pankhurst, Luke Anson, Mariateresa de Cesare, et al. Rapid antibiotic-resistance predictions from genome sequence data for staphylococcus aureus and mycobacterium tuberculosis. *Nature Communications*, 6(1) : 1–10, 2015.

[23] Priscille Brodin, Karin Eiglmeier, Magali Marmiesse, Alain Billault, Thierry Garnier, Stefan Niemann, Stewart T Cole, and Roland Brosch. Bacterial artificial chromosome-based comparative genomic analysis identifies mycobacterium microti as a natural esat-6 deletion mutant. *Infection and immunity*, 70 :5568–78, 10 2002. ISSN 0019-9567. doi : 10.1128/IAI.70.10.5568-5578.2002.

[24] R. Brosch, S. V. Gordon, M. Marmiesse, P. Brodin, C. Buchrieser, K. Eiglmeier, T. Garnier, C. Gutierrez, G. Hewinson, K. Kremer, L. M. Parsons, A. S. Pym, S. Samper, D. van Soolingen, and S. T. Cole. A new evolutionary scenario for the Mycobacterium tuberculosis complex. *Proceedings of the National Academy of Sciences*, 99(6) :3684–3689, mar 2002. ISSN 0027-8424. doi : 10.1073/pnas.052548299. URL http://www.pnas.org/lookup/doi/10.1073/pnas.052548299.

[25] Roland Brosch, Stephen V. Gordon, Alain Billault, Thierry Garnier, Karin Eiglmeier, Catherine Soravito, Bart G. Barrell, and Stewart T. Cole. Use of a Mycobacterium tuberculosisH37Rv Bacterial Artificial Chromosome Library for Genome Mapping, Sequencing, and Comparative Genomics. *Infection and Immunity*, 66(5) : 2221–2229, may 1998. ISSN 1098-5522. doi : 10.1128/IAI.66.5.2221-2229.1998. URL https://iai.asm.org/content/66/5/2221.

[26] Roland Brosch, Stephen V. Gordon, Carmen Buchrieser, Alexander S. Pym, Thierry Garnier, and Stewart T. Cole. Comparative genomics uncovers large tandem chromosomal duplications in mycobacterium bovis bcg pasteur. *Yeast*, 1 :111–123, 2000. ISSN 0749-503X. doi : 10.1002/1097-0061(20000630)17:2<111::AID-YEA17>3.0.CO;2-G.

[27] T. A. Brown. *Genomes 4*. Garland Science, Other titles : Genomes | Genomes four Description : 4th. | New York, NY : Garland Science, [2017] | Preceded by :, dec 2018. ISBN 9781315226828. doi : 10.1201/9781315226828. URL https://www.taylorfrancis.com/books/9781351851299.

[28] Stephen J Bush, Dona Foster, David W Eyre, Emily L Clark, Nicola De Maio, Liam P Shaw, Nicole Stoesser, Tim E A Peto, Derrick W Crook, and A Sarah Walker. Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism–calling pipelines. *GigaScience*, 9(2), feb 2020. ISSN 2047-217X. doi : 10.1093/gigascience/giaa007. URL https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giaa007/5728470.

[29] Antonio Campos-Neto. Mycobacterium tuberculosis strain 18b, a useful non-virulent streptomycin dependent mutant to study latent tuberculosis as well as for in vivo and in vitro testing of anti-tuberculosis drugs. *Tuberculosis*, 99 :54–55, 7 2016. ISSN 14729792. doi : 10.1016/j.tube.2016.04.006.

[30] Nicola Casali, Vladyslav Nikolayevskyy, Yanina Balabanova, Olga Ignatyeva, Irina Kontsevaya, Simon R Harris, Stephen D Bentley, Julian Parkhill, Sergey Nejentsev, Sven E Hoffner, Rolf D Horstmann, Timothy Brown, and Francis Drobniewski. Microevolution of extensively drug-resistant tuberculosis in russia. *Genome Res.*, 22 (4) :735–745, April 2012.

[31] Nicola Casali, Vladyslav Nikolayevskyy, Yanina Balabanova, Simon R Harris, Olga Ignatyeva, Irina Kontsevaya, Jukka Corander, Josephine Bryant, Julian Parkhill, Sergey Nejentsev, Rolf D Horstmann, Timothy Brown, and Francis Drobniewski. Evolution and transmission of drug-resistant tuberculosis in a russian population. *Nat. Genet.*, 46(3) :279–286, March 2014.

[32] Center for Substance Abuse Treatment. *The tuberculosis epidemic : Legal and ethical issues for alcohol and other drug treatment providers*. Substance Abuse and Mental Health Services Administration (US), Rockville (MD), 1995.

[33] Shifu Chen, Yanqing Zhou, Yaru Chen, and Jia Gu. fastp : an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17) :i884–i890, sep 2018. ISSN 1367-4803. doi : 10.1093/bioinformatics/bty560. URL https://academic.oup.com/bioinformatics/article/34/17/i884/5093234.

[34] Wenjing Chen, Tapan Biswas, Vanessa R Porter, Oleg V Tsodikov, and Sylvie Garneau-Tsodikova. Unusual regioversatility of acetyltransferase eis, a cause of drug resistance in xdr-tb. *Proceedings of the National Academy of Sciences*, 108(24) : 9804–9808, 2011.

[35] Yiwang Chen, Qi Jiang, Qingyun Liu, Mingyu Gan, Howard E Takiff, and Qian Gao. Whole-genome sequencing exhibits better diagnostic performance than variable-number tandem repeats for identifying mixed infections of mycobacterium tuberculosis. *Microbiology Spectrum*, pages e03570–22, 2023.

[36] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena : An open platform for evaluating llms by human preference. 3 2024.

[37] Pablo Cingolani, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J Land, Xiangyi Lu, and Douglas M Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff : Snps in the genome of drosophila melanogaster strain w1118 ; iso-2 ; iso-3. *Fly*, 6 :80–92. ISSN 1933-6942. doi : 10.4161/fly.19695.

[38] Daniela M. Cirillo, Paolo Miotto, and Enrico Tortoli. *Evolution of Phenotypic and Molecular Drug Susceptibility Testing*, pages 221–246. 2017. doi : 10.1007/978-3-319-64371-7_12.

[39] Adam Cohen, Victor Dahl Mathiasen, Thomas Schön, and Christian Wejse. The global prevalence of latent tuberculosis : a systematic review and meta-analysis. *European Respiratory Journal*, 54 :1900655, 9 2019. ISSN 0903-1936. doi : 10.1183/13993003.00655-2019.

[40] Eugene C. Cole and Carl E. Cook. Characterization of infectious aerosols in health care facilities : An aid to effective engineering controls and preventive strategies. *American Journal of Infection Control*, 26 :453–464, 8 1998. ISSN 01966553. doi : 10.1016/S0196-6553(98)70046-X.

[41] S. T. Cole, R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S. V. Gordon, K. Eiglmeier, S. Gas, C. E. Barry, F. Tekaia, K. Badcock, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. Davies, K. Devlin, T. Feltwell, S. Gentles, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels, A. Krogh, J. McLean, S. Moule, L. Murphy, K. Oliver, J. Osborne, M. A. Quail, M.-A. Rajandream, J. Rogers, S. Rutter, K. Seeger, J. Skelton, R. Squares, S. Squares, J. E. Sulston, K. Taylor, S. Whitehead, and B. G. Barrell. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature*, 393(6685) :537–544, jun 1998. ISSN 0028-0836. doi : 10.1038/31159. URL http://www.nature.com/articles/31159.

[42] F. Coll, R. McNerney, M. D. Preston, J. A. Guerra-Assunção, A. Warry, G. Hill-Cawthorne, K. Mallard, M. Nair, A. Miranda, A. Alves, J. Perdigão, M. Viveiros, I. Portugal, Z. Hasan, R. Hasan, J. R. Glynn, N. Martin, A. Pain, and T. G. Clark. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. *Genome medicine*, 7(1) :1–11, 2015.

[43] Francesc Coll, Ruth McNerney, José Afonso Guerra-Assunção, Judith R Glynn, João Perdigão, Miguel Viveiros, Isabel Portugal, Arnab Pain, Nigel Martin, and Taane G Clark. A robust SNP barcode for typing mycobacterium tuberculosis complex strains. *Nat. Commun.*, 5(1), September 2014.

[44] Francesc Coll, Ruth McNerney, José Afonso Guerra-Assunção, Judith R Glynn, João Perdigão, Miguel Viveiros, Isabel Portugal, Arnab Pain, Nigel Martin, and Taane G Clark. A robust snp barcode for typing mycobacterium tuberculosis complex strains. *Nature communications*, 5(1) :1–5, 2014.

[45] Francesc Coll, Mark Preston, José Afonso Guerra-Assunção, Grant Hill-Cawthorn, David Harris, João Perdigão, Miguel Viveiros, Isabel Portugal, Francis Drobniewski, Sebastien Gagneux, Judith R. Glynn, Arnab Pain, Julian Parkhill, Ruth McNerney, Nigel Martin, and Taane G. Clark. Polytb : A genomic variation map for mycobacterium tuberculosis. *Tuberculosis*, 94 :346–354, 5 2014. ISSN 14729792. doi : 10.1016/j.tube.2014.02.005.

[46] Iñaki Comas, Jaidip Chakravartti, Peter M Small, James Galagan, Stefan Niemann, Kristin Kremer, Joel D Ernst, and Sebastien Gagneux. Human t cell epitopes of mycobacterium tuberculosis are evolutionarily hyperconserved. *Nature genetics*, 42 (6) :498–503, 2010.

[47] Iñaki Comas, Sonia Borrell, Andreas Roetzer, Graham Rose, Bijaya Malla, Midori Kato-Maeda, James Galagan, Stefan Niemann, and Sebastien Gagneux. Whole-genome sequencing of rifampicin-resistant mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes. *Nat. Genet.*, 44(1) :106–110, December 2011.

[48] Iñaki Comas, Mireia Coscolla, Tao Luo, Sonia Borrell, Kathryn E Holt, Midori Kato-Maeda, Julian Parkhill, Bijaya Malla, Stefan Berg, Guy Thwaites, Dorothy Yeboah-Manu, Graham Bothamley, Jian Mei, Lanhai Wei, Stephen Bentley, Simon R Harris, Stefan Niemann, Roland Diel, Abraham Aseffa, Qian Gao,

Douglas Young, and Sebastien Gagneux. Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. *Nature Genetics*, 45(10) :1176–1182, oct 2013. ISSN 1061-4036. doi : 10.1038/ng.2744. URL http://www.nature.com/articles/ng.2744https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3800747/.

[49] Iñaki Comas. *Genomic Epidemiology of Tuberculosis*, pages 79–93. 2017. doi : 10.1007/978-3-319-64371-7_4.

[50] CNGb Consortium. Cngb pathogen variation database. https://db.cngb.org/pvd/global-result/?global=tuberculosis, 2021.

[51] Enterobase Consortium. Enterobase : Bacterial genomic data and analysis. https://enterobase.warwick.ac.uk/species/index/mycobacterium, 2017.

[52] PathogenSeq Consortium. Pathogenseq : Tools for bacterial genomic analysis. http://pathogenseq.lshtm.ac.uk/#portfolioTB6, 2017.

[53] TBDB Consortium. Tbdb : Tuberculosis database. http://www.tbdb.org/, 2013.

[54] TubercuList Consortium. Tuberculist : Mycobacterium tuberculosis h37rv database. http://tuberculist.epfl.ch/, 2010.

[55] Patricia Constant, Esther Perez, Wladimir Malaga, Marie-Antoinette Lanéelle, Olivier Saurel, Mamadou Daffé, and Christophe Guilhot. Role of the pks15/1 gene in the biosynthesis of phenolglycolipids in the mycobacterium tuberculosiscomplex. *Journal of Biological Chemistry*, 277(41) :38148–38158, 2002.

[56] Patricia Constant, Esther Perez, Wladimir Malaga, Marie-Antoinette Lanéelle, Olivier Saurel, Mamadou Daffé, and Christophe Guilhot. Role of the pks15/1 gene in the biosynthesis of phenolglycolipids in the mycobacterium tuberculosiscomplex. *Journal of Biological Chemistry*, 277 :38148–38158, 10 2002. ISSN 00219258. doi : 10.1074/jbc.M206538200.

[57] Mireia Coscolla and Sebastien Gagneux. Consequences of genomic diversity in Mycobacterium tuberculosis. *Seminars in Immunology*, 26(6) :431–444, dec 2014. ISSN 10445323. doi : 10.1016/j.smim.2014.09.012. URL https://linkinghub.elsevier.com/retrieve/pii/S1044532314000967.

[58] Mireia Coscolla, Astrid Lewin, Sonja Metzger, Kerstin Maetz-Rennsing, Sébastien Calvignac-Spencer, Andreas Nitsche, Pjotr Wojtek Dabrowski, Aleksandar Radonic, Stefan Niemann, Julian Parkhill, Emmanuel Couacy-Hymann, Julia Feldman, Iñaki Comas, Christophe Boesch, Sebastien Gagneux, and Fabian H. Leendertz. Novel mycobacterium tuberculosis complex isolate from a wild chimpanzee. *Emerging Infectious Diseases*, 19 :969–976, 6 2013. ISSN 1080-6040. doi : 10.3201/eid1906.121012.

[59] Mireia Coscolla, Daniela Brites, Fabrizio Menardo, Chloe Loiseau, Sonia Borrell, Isaac Darko Otchere, Adwoa Asante-Poku, Prince Asare, Leonor Sánchez-Busó, Florian Gehre, C N'Dira Sanoussi, Martin Antonio, Affolabi Dissou, Paula Ruiz-Rodriguez, Janet Fyfe, Erik Böttger, Patrick Becket, Stefan Niemann, Abraham Alabi, Martin Grobusch, Robin Kobbe, Julian Parkhill, Christian Beisel, Lukas Fenner, Conor Meehan, Simon Harris, Bouke De Jong, Dorothy Yeboah-Manu, and Sebastien Gagneux. Phylogenomics of Mycobacterium africanum reveals a new lineage and a complex evolutionary history, 2020. URL http://europepmc.org/abstract/PPR/PPR174375https://doi.org/10.1101/2020.06.10.141788.

[60] Mireia Coscolla, Sebastien Gagneux, Fabrizio Menardo, Chloé Loiseau, Paula Ruiz-Rodriguez, Sonia Borrell, Isaac Darko Otchere, Adwoa Asante-Poku, Prince Asare, Leonor Sánchez-Busó, et al. Phylogenomics of mycobacterium africanum reveals a new lineage and a complex evolutionary history. *Microbial genomics*, 7(2), 2021.

[61] Mireia Coscolla, Sebastien Gagneux, Fabrizio Menardo, Chloé Loiseau, Paula Ruiz-Rodriguez, Sonia Borrell, Isaac Darko Otchere, Adwoa Asante-Poku, Prince Asare, Leonor Sánchez-Busó, Florian Gehre, C. N'Dira Sanoussi, Martin Antonio, Dissou Affolabi, Janet Fyfe, Patrick Beckert, Stefan Niemann, Abraham S. Alabi, Martin P. Grobusch, Robin Kobbe, Julian Parkhill, Christian Beisel, Lukas Fenner, Erik C. Böttger, Conor J. Meehan, Simon R. Harris, Bouke C. de Jong, Dorothy Yeboah-Manu, and Daniela Brites. Phylogenomics of mycobacterium africanum reveals a new lineage and a complex evolutionary history. *Microbial Genomics*, 7, 2 2021. ISSN 2057-5858. doi : 10.1099/mgen.0.000477.

[62] Mireia Coscolla, Sebastien Gagneux, Fabrizio Menardo, Chloé Loiseau, Paula Ruiz-Rodriguez, Sonia Borrell, Isaac Darko Otchere, Adwoa Asante-Poku, Prince Asare, Leonor Sánchez-Busó, Florian Gehre, C. N'Dira Sanoussi, Martin Antonio, Dissou Affolabi, Janet Fyfe, Patrick Beckert, Stefan Niemann, Abraham S. Alabi, Martin P. Grobusch, Robin Kobbe, Julian Parkhill, Christian Beisel, Lukas Fenner, Erik C. Böttger, Conor J. Meehan, Simon R. Harris, Bouke C. de Jong, Dorothy Yeboah-Manu, and Daniela Brites. Phylogenomics of mycobacterium africanum reveals a new lineage and a complex evolutionary history. *Microbial Genomics*, 7(2) :000477, 2021. ISSN 2057-5858. doi : https://doi.org/10.1099/mgen.0.000477. URL https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000477.

[63] David Couvin, Audrey David, Thierry Zozio, and Nalin Rastogi. Macro-geographical specificities of the prevailing tuberculosis epidemic as seen through SITVIT2, an updated version of the Mycobacterium tuberculosis genotyping database. *Infection, Genetics and Evolution*, 72 :31–43, aug 2019. ISSN 15671348. doi : 10.1016/j.meegid.2018.12.030. URL https://linkinghub.elsevier.com/retrieve/pii/S1567134818309699.

[64] David Couvin, Audrey David, Thierry Zozio, and Nalin Rastogi. Macro-geographical specificities of the prevailing tuberculosis epidemic as seen through SITVIT2, an updated version of the mycobacterium tuberculosis genotyping database. *Infect. Genet. Evol.*, 72 :31–43, August 2019.

[65] David Couvin, Audrey David, Thierry Zozio, and Nalin Rastogi. Macro-geographical specificities of the prevailing tuberculosis epidemic as seen through sitvit2, an updated version of the mycobacterium tuberculosis genotyping database. *Infection, Genetics and Evolution*, 72 :31–43, 2019.

[66] David Couvin, Erick Stattner, Wilfried Segretier, Damien Cazenave, and Nalin Rastogi. simpitb – a pipeline designed to extract meaningful information from whole genome sequencing data of mycobacterium tuberculosis complex, allows to combine genomic, phylogenetic and clustering analyses in existing sitvit databases. *Infection, Genetics and Evolution*, 113 :105466, 2023. ISSN 1567-1348. doi : https://doi.org/10.1016/j.meegid.2023.105466. URL https://www.sciencedirect.com/science/article/pii/S1567134823000643.

[67] Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, and Heng Li. Twelve years of samtools and bcftools. *GigaScience*, 10, 1 2021. ISSN 2047-217X. doi : 10.1093/gigascience/giab008.

[68] Thomas M. Daniel. The history of tuberculosis. *Respiratory Medicine*, 100(11) : 1862–1870, nov 2006. ISSN 09546111. doi : 10.1016/j.rmed.2006.08.006. URL https://linkinghub.elsevier.com/retrieve/pii/S095461110600401X.

[69] S. Das, C.N. Paramasivan, D.B. Lowrie, R. Prabhakar, and P.R. Narayanan. IS6110 restriction fragment length polymorphism typing of clinical isolates of Mycobacterium tuberculosis from patients with pulmonary tuberculosis in Madras, South India. *Tubercle and Lung Disease*, 76(6) :550–554, dec 1995. ISSN 09628479. doi : 10.1016/0962-8479(95)90533-2. URL https://linkinghub.elsevier.com/retrieve/pii/0962847995905332.

[70] Bouke C. de Jong, Martin Antonio, and Sebastien Gagneux. Mycobacterium africanum—Review of an Important Cause of Human Tuberculosis in West Africa. *PLoS Neglected Tropical Diseases*, 4(9) :e744, sep 2010. ISSN 1935-2735. doi : 10.1371/journal.pntd.0000744. URL https://dx.plos.org/10.1371/journal.pntd.0000744.

[71] M de Vos, B Müller, S Borrell, P A Black, P D van Helden, R M Warren, S Gagneux, and T C Victor. Putative compensatory mutations in the rpoc gene of rifampin-resistant mycobacterium tuberculosis are associated with ongoing transmission. *Antimicrob. Agents Chemother.*, 57(2) :827–832, February 2013.

[72] Christophe Demay, Benjamin Liens, Thomas Burguière, Véronique Hill, David Couvin, Julie Millet, Igor Mokrousov, Christophe Sola, Thierry Zozio, and Nalin Rastogi. SITVITWEB – A publicly available international multimarker database for studying Mycobacterium tuberculosis genetic diversity and molecular epidemiology. *Infection, Genetics and Evolution*, 12(4) :755–766, jun 2012. ISSN 15671348. doi : 10.1016/j.meegid.2012.02.004. URL https://linkinghub.elsevier.com/retrieve/pii/S1567134812000317.

[73] Johan T den Dunnen, Raymond Dalgleish, Donna R Maglott, Reece K Hart, Marc S Greenblatt, Jean McGowan-Jordan, Anne-Francoise Roux, Timothy Smith, Stylianos E Antonarakis, and Peter E M Taschner. Hgvs recommendations for the description of sequence variants : 2016 update. *Human mutation*, 37 :564–9, 6 2016. ISSN 1098-1004. doi : 10.1002/humu.22981.

[74] Isabelle Devaux, Kristin Kremer, Herre Heersma, and Dick Van Soolingen. Clusters of multidrug-resistant mycobacterium tuberculosis cases, europe. *Emerg. Infect. Dis.*, 15(7) :1052–1060, July 2009.

[75] Anzaan Dippenaar, Sven David Charles Parsons, Samantha Leigh Sampson, Ruben Gerhard van der Merwe, Julian Ashley Drewe, Abdallah Musa Abdallah, Kabengele Keith Siame, Nicolaas Claudius Gey van Pittius, Paul David van Helden, Arnab Pain, and Robin Mark Warren. Whole genome sequence analysis of mycobacterium suricattae. *Tuberculosis*, 95 :682–688, 12 2015. ISSN 14729792. doi : 10.1016/j.tube.2015.10.001.

[76] Maxwell Droznin, Allen Johnson, and Asal Mohamadi Johnson. Multidrug resistant tuberculosis in prisons located in former soviet countries : A systematic review. *PLoS One*, 12(3) :e0174373, March 2017.

[77] Pierre Dupuy, Shreya Ghosh, Oyindamola Adefisayo, John Buglino, Stewart Shuman, and Michael S Glickman. Distinctive roles of translesion polymerases dinb1 and dnae2 in diversification of the mycobacterial genome through substitution and frameshift mutagenesis. *Nature Communications*, 13(1) :4493, 2022.

[78] Mark T. W. Ebbert, Mark E. Wadsworth, Lyndsay A. Staley, Kaitlyn L. Hoyt, Brandon Pickett, Justin Miller, John Duce, John S. K. Kauwe, and Perry G. Ridge. Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC Bioinformatics*, 17(S7) :239, jul 2016. ISSN 1471-2105. doi : 10.1186/s12859-016-1097-3. URL http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1097-3.

[79] Karen Eilbeck, Suzanna E Lewis, Christopher J Mungall, Mark Yandell, Lincoln Stein, Richard Durbin, and Michael Ashburner. The sequence ontology : a tool for the unification of genome annotations. *Genome biology*, 6 :R44, 2005. ISSN 1474-760X. doi : 10.1186/gb-2005-6-5-r44.

[80] Joel D. Ernst. The immunological life cycle of tuberculosis. *Nature Reviews Immunology*, 12(8) :581–591, aug 2012. ISSN 1474-1733. doi : 10.1038/nri3259. URL http://www.nature.com/articles/nri3259.

[81] Kiatichai Faksri, Eryu Xia, Jun Hao Tan, Yik-Ying Teo, and Rick Twee-Hee Ong. In silico region of difference (RD) analysis of Mycobacterium tuberculosis complex from sequence reads using RD-Analyzer. *BMC Genomics*, 17(1) :847, dec 2016. ISSN 1471-2164. doi : 10.1186/s12864-016-3213-1. URL http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-016-3213-1.

[82] Xingzhun Fan. Zhongguo bingshi xinyi (new meaning of disease in chinese history). *Beijing : Zhongyi guji chuban.[Google Scholar]*, 1989.

[83] Joseph Felsenstein. *Inferring Phylogenies*. 2004.

[84] J Ferrer. Pleural tuberculosis. *The European respiratory journal*, 10 :942–7, 4 1997. ISSN 0903-1936.

[85] S. Feuerriegel, V. Schleusener, P. Beckert, T. A. Kohl, P. Miotto, D. M. Cirillo, A. M. Cabibbe, S. Niemann, and K. Fellenberg. Phyresse : a web tool delineating mycobacterium tuberculosis antibiotic resistance and lineage from whole-genome sequencing data. *Journal of clinical microbiology*, 53(6) :1908–1914, 2015.

[86] Christopher B Ford, Rupal R Shah, Midori Kato Maeda, Sebastien Gagneux, Megan B Murray, Ted Cohen, James C Johnston, Jennifer Gardy, Marc Lipsitch, and Sarah M Fortune. Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nat. Genet.*, 45(7) :784–790, July 2013.

[87] Luca Freschi, Roger Vargas, Ashaque Husain, SM Kamal, Alena Skrahina, Sabira Tahseen, Nazir Ismail, Anna Barbova, Stefan Niemann, Daniela Maria Cirillo, et al. Population structure, biogeography and transmissibility of mycobacterium tuberculosis. *Nature communications*, 12(1) :1–11, 2021.

[88] Luca Freschi, Roger Vargas, Jr, Ashaque Husain, S M Mostofa Kamal, Alena Skrahina, Sabira Tahseen, Nazir Ismail, Anna Barbova, Stefan Niemann, Daniela Maria Cirillo, Anna S Dean, Matteo Zignol, and Maha Reda Farhat. Population structure, biogeography and transmissibility of mycobacterium tuberculosis. *Nat. Commun.*, 12(1) :6099, October 2021.

[89] John A. Frith. History of tuberculosis. part 1 - phthisis, consumption and the white plague. *Journal of Military and Veterans' Health*, 22 :29–35, 2014.

[90] Richard Frothingham and Winifred A. Meeker-O'Connell. Genetic diversity in the mycobacterium tuberculosis complex based on variable numbers of tandem dna repeats. *Microbiology*, 144 :1189–1196, 5 1998. ISSN 1350-0872. doi : 10.1099/00221287-144-5-1189.

[91] S. Gagneux, K. DeRiemer, T. Van, M. Kato-Maeda, B. C. de Jong, S. Narayanan, M. Nicol, S. Niemann, K. Kremer, M. C. Gutierrez, M. Hilty, P. C. Hopewell, and P. M. Small. Variable host-pathogen compatibility in Mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences*, 103(8) :2869–2873, feb 2006. ISSN 0027-8424. doi : 10.1073/pnas.0511240103. URL http://www.pnas.org/cgi/doi/10.1073/pnas.0511240103.

[92] Sebastien Gagneux. Ecology and evolution of mycobacterium tuberculosis. *Nature Reviews Microbiology*, 16 :202–213, 4 2018. ISSN 1740-1526. doi : 10.1038/nrmicro.2018.8.

[93] Sebastien Gagneux. Ecology and evolution of mycobacterium tuberculosis. *Nature Reviews Microbiology*, 16(4) :202–213, 2018.

[94] Sebastien Gagneux and Peter M Small. Global phylogeography of mycobacterium tuberculosis and implications for tuberculosis product development. *The Lancet Infectious Diseases*, 7 :328–337, 5 2007. ISSN 14733099. doi : 10.1016/S1473-3099(07)70108-1.

[95] M.Carmen Garcia-Pelayo, Karina C Caimi, Jacqueline K Inwald, Jason Hinds, Fabiana Bigi, Maria I Romano, Dick van Soolingen, R.Glyn Hewinson, Angel Cataldi, and Stephen V Gordon. Microarray analysis of mycobacterium microti reveals deletion of genes encoding pe-ppe proteins and esat-6 family antigens. *Tuberculosis*, 84 : 159–166, 1 2004. ISSN 14729792. doi : 10.1016/j.tube.2003.12.002.

[96] Jennifer L Gardy, James C Johnston, Shannan J Ho Sui, Victoria J Cook, Lena Shah, Elizabeth Brodkin, Shirley Rempel, Richard Moore, Yongjun Zhao, Robert Holt, et al. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *New England Journal of Medicine*, 364(8) :730–739, 2011.

[97] Thierry Garnier, Karin Eiglmeier, Jean-Christophe Camus, Nadine Medina, Huma Mansoor, Melinda Pryor, Stephanie Duthoy, Sophie Grondin, Celine Lacroix, Christel Monsempe, Sylvie Simon, Barbara Harris, Rebecca Atkin, Jon Doggett, Rebecca Mayes, Lisa Keating, Paul R. Wheeler, Julian Parkhill, Bart G. Barrell, Stewart T. Cole, Stephen V. Gordon, and R. Glyn Hewinson. The complete genome sequence of mycobacterium bovis. *Proceedings of the National Academy of Sciences*, 100 : 7877–7882, 6 2003. ISSN 0027-8424. doi : 10.1073/pnas.1130426100.

[98] Erik Garrison and Gabor Marth. Haplotype-based variant detection from short-read sequencing. 7 2012.

[99] Marjorie P Golden and Holenarasipur R Vikram. Extrapulmonary tuberculosis : an overview. *American family physician*, 72 :1761–8, 11 2005. ISSN 0002-838X.

[100] Stephen V. Gordon, Roland Brosch, Alain Billault, Thierry Garnier, Karin Eiglmeier, and Stewart T. Cole. Identification of variable regions in the genomes of tubercle bacilli using bacterial artificial chromosome arrays. *Molecular Microbiology*, 32(3) :643–655, may 1999. ISSN 0950-382X. doi : 10.1046/j.1365-2958.1999.01383.x. URL http://doi.wiley.com/10.1046/j.1365-2958.1999.01383.x.

[101] Dan Graur. *Single-base Mutation*. Wiley, 7 2006. doi : 10.1038/npg.els.0005093.

[102] Peter M. A. Groenen, Annelies E. Bunschoten, Dick van Soolingen, and Jan D. A. van Errtbden. Nature of DNA polymorphism in the direct repeat cluster of Mycobacterium tuberculosis ; application for strain differentiation by a novel typing method. *Molecular Microbiology*, 10(5) :1057–1065, dec 1993. ISSN 0950-382X. doi : 10.1111/j.1365-2958.1993.tb00976.x. URL http://doi.wiley.com/10.1111/j.1365-2958.1993.tb00976.x.

[103] Qingtian Guan, Musa Garbati, Sara Mfarrej, Talal AlMutairi, Thomas Laval, Albel Singh, Shamsudeen Fagbo, Alicia Smyth, John A Browne, Muhammad Amin urRahman, Alya Alruwaili, Anwar Hoosen, Conor J Meehan, Chie Nakajima, Yasuhiko Suzuki, Caroline Demangel, Apoorva Bhatt, Stephen V Gordon, Faisal AlAsmari, and Arnab Pain. Insights into the ancestry evolution of the mycobacterium tuberculosis complex from analysis of mycobacterium riyadhense. *NAR Genomics and Bioinformatics*, 3, 6 2021. ISSN 2631-9268. doi : 10.1093/nargab/lqab070.

[104] José Afonso Guerra-Assunção, Rein M. G. J. Houben, Amelia C. Crampin, Themba Mzembe, Kim Mallard, Francesc Coll, Palwasha Khan, Louis Banda, Arthur Chiwaya, Rui P. A. Pereira, Ruth McNerney, David Harris, Julian Parkhill, Taane G. Clark, and Judith R. Glynn. Recurrence due to relapse or reinfection with mycobacterium tuberculosis : A whole-genome sequencing approach in a large, population-based cohort with a high hiv infection prevalence and active follow-up. *Journal of Infectious Diseases*, 211 :1154–1163, 4 2015. ISSN 0022-1899. doi : 10.1093/infdis/jiu574.

[105] Christophe Guyeux, Gaetan Senelle, Guislaine Refrégier, Emmanuelle Cambau, and Christophe Sola. Description of a new ancestral lineage of the L2 Mycobacterium tuberculosis complex in Japan and Revision of the current L2 Lineage/sublineage nomenclature. 2021.

[106] Christophe Guyeux, Christophe Sola, Camille Noûs, and Guislaine Refrégier. CRISPRbuilder-TB : "CRISPR-builder for tuberculosis". exhaustive reconstruction of the CRISPR locus in mycobacterium tuberculosis complex using SRA. *PLoS Comput. Biol.*, 17(3) :e1008500, March 2021.

[107] Christophe Guyeux, Christophe Sola, Camille Noûs, and Guislaine Refrégier. Crisprbuilder-tb : "crispr-builder for tuberculosis". exhaustive reconstruction of the crispr locus in mycobacterium tuberculosis complex using sra. *PLOS Computational Biology*, 17 :e1008500, 3 2021. ISSN 1553-7358. doi : 10.1371/journal.pcbi.1008500.

[108] Christophe Guyeux, Gaetan Senelle, Guislaine Refrégier, Florence Bretelle-Establet, Emmanuelle Cambau, and Christophe Sola. Connection between two historical tuberculosis outbreak sites in japan, honshu, by a new ancestral mycobacterium tuberculosis l2 sublineage. *Epidemiology & Infection*, 150, 2022.

[109] Christophe Guyeux, Gaëtan Senelle, Adrien Le Meur, Christophe Sola, and Guislaine Refrégier. The hidden diversity of mycobacterium tuberculosis complex in africa : the new l10 and the possible diversification histories of the complex. In *44th Annual Congress of the European Society of Mycobacteriology (ESM 2024)*, Brugges, Belgium, jun 2024.

[110] Christophe Guyeux, Gaetan Senelle, Adrien Le Meur, Philip Supply, Cyril Gaudin, Jody E. Phelan, Taane G Clark, Leen Rigouts, Bouke de Jong, Christophe Sola, and Guislaine Refrégier. Newly identified mycobacterium africanum lineage 10, central africa. *Emerging Infectious Diseases*, 30, 10 2024. ISSN 1080-6040. doi : 10.3201/eid3003.231466.

[111] Sebastian M Gygli, Chloé Loiseau, Levan Jugheli, Natia Adamia, Andrej Trauner, Miriam Reinhard, Amanda Ross, Sonia Borrell, Rusudan Aspindzelashvili, Nino Maghradze, Klaus Reither, Christian Beisel, Nestani Tukvadze, Zaza Avaliani, and Sebastien Gagneux. Prisons as ecological drivers of fitness-compensated multidrug-resistant mycobacterium tuberculosis. *Nat. Med.*, 27(7) :1171–1177, July 2021.

[112] James Hadfield, Colin Megill, Sidney M Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, and Richard A Neher. Nextstrain : real-time tracking of pathogen evolution. *Bioinformatics*, 34(23) :4121–4123, 2018.

[113] Seung Jung Han, Taeksun Song, Yong-Joon Cho, Jong-Seok Kim, Soo Young Choi, Hye-Eun Bang, Jongsik Chun, Gill-Han Bai, Sang-Nae Cho, and Sung Jae Shin. Complete genome sequence of mycobacterium tuberculosis K from a korean high school outbreak, belonging to the beijing family. *Stand. Genomic Sci.*, 10(1) :78, October 2015.

[114] P W Hermans, D van Soolingen, E M Bik, P E de Haas, J W Dale, and J D van Embden. Insertion element IS987 from Mycobacterium bovis BCG is located in a hot-spot integration region for insertion elements in Mycobacterium tuberculosis complex strains. *Infection and Immunity*, 59(8) :2695–2705, 1991. ISSN 0019-9567. doi : 10.1128/IAI.59.8.2695-2705.1991. URL https://iai.asm.org/content/59/8/2695.

[115] P W Hermans, D van Soolingen, and J D van Embden. Characterization of a major polymorphic tandem repeat in mycobacterium tuberculosis and its potential use in the epidemiology of mycobacterium kansasii and mycobacterium gordonae. *Journal of Bacteriology*, 174 :4157–4165, 6 1992. ISSN 0021-9193. doi : 10.1128/jb.174.12.4157-4165.1992.

[116] Israel Hershkovitz, Helen D. Donoghue, David E. Minnikin, Gurdyal S. Besra, Oona Y-C. Lee, Angela M. Gernaey, Ehud Galili, Vered Eshed, Charles L. Greenblatt, Eshetu Lemma, Gila Kahila Bar-Gal, and Mark Spigelman. Detection and molecular characterization of 9000-year-old mycobacterium tuberculosis from a neolithic settlement in the eastern mediterranean. *PLoS ONE*, 3 :e3426, 10 2008. ISSN 1932-6203. doi : 10.1371/journal.pone.0003426.

[117] H Herzog. History of tuberculosis. *Respiration ; international review of thoracic diseases*, 65 :5–15, 1998. ISSN 0025-7931. doi : 10.1159/000029220.

[118] Gareth Highnam, Jason J. Wang, Dean Kusler, Justin Zook, Vinaya Vijayan, Nir Leibovich, and David Mittelman. An analytical framework for optimizing variant discovery from personal genomes. *Nature Communications*, 6(1) :6275, may 2015. ISSN 2041-1723. doi : 10.1038/ncomms7275. URL http://www.nature.com/articles/ncomms7275.

[119] Hippocrates. *Of the Epidemics*. -400.

[120] A. E. Hirsh, A. G. Tsolaki, K. DeRiemer, M. W. Feldman, and P. M. Small. Stable association between strains of Mycobacterium tuberculosis and their human host populations. *Proceedings of the National Academy of Sciences*, 101(14) :4871–4876, apr 2004. ISSN 0027-8424. doi : 10.1073/pnas.0305627101. URL http://www.pnas.org/cgi/doi/10.1073/pnas.0305627101.

[121] Aaron E Hirsh, Anthony G Tsolaki, Kathryn DeRiemer, Marcus W Feldman, and Peter M Small. Stable association between strains of mycobacterium tuberculosis and their human host populations. *Proc. Natl. Acad. Sci. U. S. A.*, 101(14) :4871–4876, April 2004.

[122] J Bradley Holmes, Eric Moyer, Lon Phan, Donna Maglott, and Brandi Kattman. Spdi : data model for variants and applications at ncbi. *Bioinformatics*, 36 :1902–1907, 3 2020. ISSN 1367-4803. doi : 10.1093/bioinformatics/btz856.

[123] M. A. Horstkotte, I. Sobottka, C. K. Schewe, P. Schafer, R. Laufs, S. Rusch-Gerdes, and S. Niemann. Mycobacterium microti Llama-Type Infection Presenting as Pulmonary Tuberculosis in a Human Immunodeficiency Virus-Positive Patient. *Journal of Clinical Microbiology*, 39(1) :406–407, jan 2001. ISSN 0095-1137. doi : 10.1128/JCM.39.1.406-407.2001. URL http://jcm.asm.org/cgi/doi/10.1128/JCM.39.1.406-407.2001.

[124] M. Hsi-Yang Fritz, R. Leinonen, G. Cochrane, and E. Birney. Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Research*, 21(5) :734–740, may 2011. ISSN 1088-9051. doi : 10.1101/gr.114819.110. URL http://genome.cshlp.org/cgi/doi/10.1101/gr.114819.110.

[125] Richard C. Huard, Michel Fabre, Petra de Haas, Luiz Claudio Oliveira Lazzarini, Dick van Soolingen, Debby Cousins, and John L. Ho. Novel genetic polymorphisms that further delineate the phylogeny of the mycobacterium tuberculosis complex. *Journal of Bacteriology*, 188 :4271–4287, 6 2006. ISSN 0021-9193. doi : 10.1128/JB.01783-05.

[126] HUGO Pan-Asian SNP Consortium, Mahmood Ameen Abdulla, Ikhlak Ahmed, Anunchai Assawamakin, Jong Bhak, Samir K Brahmachari, Gayvelline C Calacal, Amit Chaurasia, Chien-Hsiun Chen, Jieming Chen, Yuan-Tsong Chen, Jiayou Chu, Eva Maria C Cutiongco-de la Paz, Maria Corazon A De Ungria, Frederick C Delfin, Juli Edo, Suthat Fuchareon, Ho Ghang, Takashi Gojobori, Junsong Han, Sheng-Feng Ho, Boon Peng Hoh, Wei Huang, Hidetoshi Inoko, Pankaj Jha, Timothy A Jinam, Li Jin, Jongsun Jung, Daoroong Kangwanpong, Jatupol Kampuansai, Giulia C Kennedy, Preeti Khurana, Hyung-Lae Kim, Kwangjoong Kim, Sangsoo Kim, Woo-Yeon Kim, Kuchan Kimm, Ryosuke Kimura, Tomohiro Koike, Supasak Kulawonganunchai, Vikrant Kumar, Poh San Lai, Jong-Young Lee, Sunghoon Lee, Edison T Liu, Partha P Majumder, Kiran Kumar Mandapati, Sangkot Marzuki, Wayne Mitchell, Mitali Mukerji, Kenji Naritomi, Chumpol Ngamphiw, Norio Niikawa, Nao Nishida, Bermseok Oh, Sangho Oh, Jun Ohashi, Akira Oka, Rick Ong, Carmencita D Padilla, Prasit Palittapongarnpim, Henry B Perdigon, Maude Elvira Phipps, Eileen Png, Yoshiyuki Sakaki, Jazelyn M Salvador, Yuliana Sandraling, Vinod Scaria, Mark Seielstad, Mohd Ros Sidek, Amit Sinha, Metawee Srikummool, Herawati Sudoyo, Sumio Sugano, Helena Suryadi, Yoshiyuki Suzuki, Kristina A Tabbada, Adrian Tan, Katsushi Tokunaga, Sissades Tongsima, Lilian P Villamor, Eric Wang, Ying Wang, Haifeng Wang, Jer-Yuarn Wu, Huasheng Xiao, Shuhua Xu, Jin Ok Yang, Yin Yao Shugart, Hyang-Sook Yoo, Wentao Yuan, Guoping Zhao, Bin Alwi Zilfalil, and Indian Genome Variation Consortium. Mapping human genetic diversity in asia. *Science*, 326 (5959) :1541–1545, December 2009.

[127] illumina. Poly-g trimming. URL https://support.illumina.com/content/dam/illumina-support/help/Illumina_DRAGEN_Bio_IT_Platform_v3_7_1000000141465/Content/SW/Informatics/Dragen/PolyG_Trimming_fDG.htm.

[128] Interchim. Blotter Systems. URL http://www.interchim.com/pp/700/blotter-systems.html.

[129] Hiroki Iwai, Masako Kato-Miyazawa, Teruo Kirikae, and Tohru Miyoshi-Akiyama. 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*, 95(6) : 843–844, December 2015. ISSN 1472-9792. doi : 10.1016/j.tube.2015.09.002. URL http://dx.doi.org/10.1016/j.tube.2015.09.002.

[130] Hiroki Iwai, Masako Kato-Miyazawa, Teruo Kirikae, and Tohru Miyoshi-Akiyama. 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 (Edinb.)*, 95(6) :843–844, December 2015.

[131] Hiroki Iwai, Masako Kato-Miyazawa, Teruo Kirikae, and Tohru Miyoshi-Akiyama. 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*, 95 :843–844, 2015.

[132] Tomotada Iwamoto, Riyo Fujiyama, Shiomi Yoshida, Takayuki Wada, Chika Shirai, and Yasuto Kawakami. Population structure dynamics of mycobacterium tuberculosis beijing strains during past decades in japan. *J. Clin. Microbiol.*, 47(10) : 3340–3343, October 2009.

[133] Tomotada Iwamoto, Louis Grandjean, Kentaro Arikawa, Noriko Nakanishi, Luz Caviedes, Jorge Coronel, Patricia Sheen, Takayuki Wada, Carmen A Taype, Marie-Anne Shaw, David A J Moore, and Robert H Gilman. Genetic diversity and transmission characteristics of beijing family strains of mycobacterium tuberculosis in peru. *PLoS One*, 7(11) :e49651, November 2012.

[134] Tomasz Jagielski, Jakko van Ingen, Nalin Rastogi, Jarosław Dziadek, Paweł K Mazur, and Jacek Bielecki. Current methods in the molecular typing of mycobacterium tuberculosis and other mycobacteria. *BioMed research international*, 2014 :645802, 2014. ISSN 2314-6141. doi : 10.1155/2014/645802.

[135] Ruud. Jansen, Jan. D. A. van Embden, Wim. Gaastra, and Leo. M. Schouls. Identification of genes that are associated with DNA repeats in prokaryotes. *Molecular Microbiology*, 43(6) :1565–1575, mar 2002. ISSN 0950-382X. doi : 10.1046/j.1365-2958.2002.02839.x. URL http://doi.wiley.com/10.1046/j.1365-2958.2002.02839.x.

[136] Alec J. Jeffreys, Annette MacLeod, Keiji Tamaki, David L. Neil, and Darren G. Monckton. Minisatellite repeat coding as a digital approach to DNA typing. *Nature*, 354(6350) :204–209, nov 1991. ISSN 0028-0836. doi : 10.1038/354204a0. URL http://www.nature.com/articles/354204a0.

[137] James C. Johnston, Ryan Cooper, and Dick Menzies. Chapter 5 : Treatment of tuberculosis disease. *Canadian Journal of Respiratory, Critical Care, and Sleep Medicine*, 6 :66–76, 3 2022. ISSN 2474-5332. doi : 10.1080/24745332.2022.2036504.

[138] J Kamerbeek, L Schouls, A Kolk, M van Agterveld, D van Soolingen, S Kuijper, A Bunschoten, H Molhuizen, R Shaw, M Goyal, and J van Embden. Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. *Journal of clinical microbiology*, 35(4) :907–914, 1997. ISSN 0095-1137. doi : 10.1128/JCM.35.4.907-914.1997. URL https://jcm.asm.org/content/35/4/907.

[139] J Kamerbeek, L Schouls, A Kolk, M van Agterveld, D van Soolingen, S Kuijper, A Bunschoten, H Molhuizen, R Shaw, M Goyal, and J van Embden. Simultaneous detection and strain differentiation of mycobacterium tuberculosis for diagnosis and epidemiology. *J. Clin. Microbiol.*, 35(4) :907–914, April 1997.

[140] Hee Yoon Kang, Takayuki Wada, Tomotada Iwamoto, Shinji Maeda, Yoshiro Murase, Seiya Kato, Hee Jin Kim, and Young Kil Park. Phylogeographical particularity

of the mycobacterium tuberculosis beijing family in south korea based on international comparison with surrounding countries. *J. Med. Microbiol.*, 59(Pt 10) : 1191–1197, October 2010.

[141] Adamandia Kapopoulou, Jocelyne M Lew, and Stewart T Cole. Mycobrowser : A web-based resource for tuberculosis research. *Nucleic acids research*, 39(suppl_1) : D633–D636, 2011.

[142] Malancha Karmakar, James M Trauer, David B Ascher, and Justin T Denholm. Hyper transmission of beijing lineage mycobacterium tuberculosis : Systematic review and meta-analysis. *J. Infect.*, 79(6) :572–581, December 2019.

[143] Peter D Karp, Daniel Weaver, Mario Latendresse, Andreas Dräger, and Suzanne Paley. Biocyc : a reference collection of pathway/genome databases. *Nucleic acids research*, 49(D1) :D743–D750, 2021.

[144] M Kato-Maeda, J T Rhee, T R Gingeras, H Salamon, J Drenkow, N Smittipat, and P M Small. Comparing genomes within the species mycobacterium tuberculosis. *Genome research*, 11 :547–54, 4 2001. ISSN 1088-9051. doi : 10.1101/gr.166401.

[145] M Kato-Maeda, S Gagneux, L L Flores, E Y Kim, P M Small, E P Desmond, and P C Hopewell. Strain classification of mycobacterium tuberculosis : congruence between large sequence polymorphisms and spoligotypes. *The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease*, 15 :131–3, 1 2011. ISSN 1815-7920.

[146] Gemma L. Kay, Martin J. Sergeant, Zhemin Zhou, Jacqueline Z.-M. Chan, Andrew Millard, Joshua Quick, Ildikó Szikossy, Ildikó Pap, Mark Spigelman, Nicholas J. Loman, Mark Achtman, Helen D. Donoghue, and Mark J. Pallen. Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in europe. *Nature Communications*, 6 :6717, 4 2015. ISSN 2041-1723. doi : 10.1038/ncomms7717.

[147] Michel Kaswa Kayomo, Vital Nkake Mbula, Muriel Aloni, Emmanuel André, Leen Rigouts, Fairouz Boutachkourt, Bouke C de Jong, Nicolas M Nkiere, and Anna S Dean. Targeted next-generation sequencing of sputum for diagnosis of drug-resistant tb : results of a national survey in democratic republic of the congo. *Scientific reports*, 10(1) :10786, 2020.

[148] Teemu Kivioja, Anna Vähärautio, Kasper Karlsson, Martin Bonke, Martin Enge, Sten Linnarsson, and Jussi Taipale. Counting absolute numbers of molecules using unique molecular identifiers. *Nature methods*, 9(1) :72–74, 2012.

[149] Tobias H Kloepper and Daniel H Huson. Drawing explicit phylogenetic networks and their integration into SplitsTree. *BMC Evol. Biol.*, 8 :22, January 2008.

[150] B J Klotoe, S Kacimi, E Costa-Conceicão, H M Gomes, R B Barcellos, S Panaiotov, D Haj Slimene, N Sikhayeva, S Sengstake, A R Schuitema, M Akhalaia, A Alenova, E Zholdybayeva, P Tarlykov, R Anthony, G Refrégier, and C Sola. Genomic characterization of MDR/XDR-TB in kazakhstan by a combination of high-throughput methods predominantly shows the ongoing transmission of L2/Beijing 94-32 central Asian/Russian clusters. *BMC Infect. Dis.*, 19(1) :553, June 2019.

[151] R. Kozak and M.A. Behr. Divergence of immunologic and protective responses of different BCG strains in a murine model. *Vaccine*, 29(7) :1519–1526, feb 2011. ISSN 0264410X. doi : 10.1016/j.vaccine.2010.12.012. URL https://linkinghub.elsevier.com/retrieve/pii/S0264410X10017500.

[152] Jocelyn E. Krebs, Elliott S. Goldstein, and Stephen T. Kilpatrick. *Lewin's Genes X*. Jones & Bartlett Learning, 2009. ISBN 0763766321. URL http://www.amazon.com/Lewins-Genes-X-Jocelyn-Krebs/dp/0763766321.

[153] Kubernetes. Kubernetes. URL https://kubernetes.io/fr/.

[154] Mikael Kubista, José Manuel Andrade, Martin Bengtsson, Amin Forootan, Jiri Jonák, Kristina Lind, Radek Sindelka, Robert Sjöback, Björn Sjögreen, Linda Strömbom, Anders Ståhlberg, and Neven Zoric. The real-time polymerase chain reaction. *Molecular Aspects of Medicine*, 27(2-3) :95–125, apr 2006. ISSN 00982997. doi : 10.1016/j.mam.2005.12.007. URL https://linkinghub.elsevier.com/retrieve/pii/S0098299705000907.

[155] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions : A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7 :453–466, 11 2019. ISSN 2307-387X. doi : 10.1162/tacl_a_00276.

[156] Luiz Claudio Oliveira Lazzarini, Richard C Huard, Neio L Boechat, Harrison M Gomes, Maranibia C Oelemann, Natalia Kurepina, Elena Shashkina, Fernanda C Q Mello, Andrea L Gibson, Milena J Virginio, Ana Grazia Marsico, W Ray Butler, Barry N Kreiswirth, Philip N Suffys, Jose Roberto Lapa E Silva, and John L Ho. Discovery of a novel mycobacterium tuberculosis lineage that is a major cause of tuberculosis in rio de janeiro, brazil. *Journal of clinical microbiology*, 45 :3891–902, 12 2007. ISSN 0095-1137. doi : 10.1128/JCM.01394-07.

[157] C. Ledergerber and C. Dessimoz. Base-calling for next-generation sequencing platforms. *Briefings in Bioinformatics*, 12 :489–497, 9 2011. ISSN 1467-5463. doi : 10.1093/bib/bbq077.

[158] Oona Y-C. Lee, Houdini H. T. Wu, Helen D. Donoghue, Mark Spigelman, Charles L. Greenblatt, Ian D. Bull, Bruce M. Rothschild, Larry D. Martin, David E. Minnikin, and Gurdyal S. Besra. Mycobacterium tuberculosis complex lipid virulence factors preserved in the 17,000-year-old skeleton of an extinct bison, bison antiquus. *PLoS ONE*, 7 :e41923, 7 2012. ISSN 1932-6203. doi : 10.1371/journal.pone.0041923.

[159] R. Leinonen, H. Sugawara, and M. Shumway. The sequence read archive. *Nucleic Acids Research*, 39 :D19–D21, 1 2011. ISSN 0305-1048. doi : 10.1093/nar/gkq1019.

[160] Jocelyne M Lew, Adamandia Kapopoulou, Louis M Jones, and Stewart T Cole. Tuberculist–10 years after. *Tuberculosis (Edinburgh, Scotland)*, 91 :1–7, 1 2011. ISSN 1873-281X. doi : 10.1016/j.tube.2010.09.008.

[161] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16) :2078–2079, aug 2009. ISSN 1367-4803. doi : 10.1093/bioinformatics/btp352. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp352.

[162] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem, 2013.

[163] Qin-Jing Li, Wei-Wei Jiao, Qing-Qin Yin, Fang Xu, Jie-Qiong Li, Lin Sun, Jing Xiao, Ying-Jia Li, Igor Mokrousov, Hai-Rong Huang, and A-Dong Shen. Compensatory mutations of rifampin resistance are associated with transmission of multidrug-resistant mycobacterium tuberculosis beijing genotype strains in china. *Antimicrob. Agents Chemother.*, 60(5) :2807–2812, May 2016.

[164] Qiao Liu, Beibei Qiu, Guoli Li, Tingting Yang, Bilin Tao, Leonardo Martinez, Limei Zhu, Jianming Wang, Xuhua Mao, and Wei Lu. Tuberculosis reinfection and relapse in eastern china : a prospective study using whole-genome sequencing. *Clinical Microbiology and Infection*, 28 :1458–1464, 11 2022. ISSN 1198743X. doi : 10.1016/j.cmi.2022.05.019.

[165] Qingyun Liu, Tao Luo, Xinran Dong, Gang Sun, Zhu Liu, Mingyun Gan, Jie Wu, Xin Shen, and Qian Gao. Genetic features of mycobacterium tuberculosis modern beijing sublineage. *Emerg. Microbes Infect.*, 5(1) :e14, February 2016.

[166] Qingyun Liu, Aijing Ma, Lanhai Wei, Yu Pang, Beibei Wu, Tao Luo, Yang Zhou, Hong-Xiang Zheng, Qi Jiang, Mingyu Gan, Tianyu Zuo, Mei Liu, Chongguang Yang, Li Jin, Iñaki Comas, Sebastien Gagneux, Yanlin Zhao, Caitlin S Pepperell, and Qian Gao. China's tuberculosis epidemic stems from historical expansion of four strains of mycobacterium tuberculosis. *Nat. Ecol. Evol.*, 2(12) :1982–1992, December 2018.

[167] Yongchao Liu, Bernt Popp, and Bertil Schmidt. CUSHAW3 : Sensitive and Accurate Base-Space and Color-Space Short-Read Alignment with Hybrid Seeding. *PLoS ONE*, 9(1) :e86869, jan 2014. ISSN 1932-6203. doi : 10.1371/journal.pone.0086869. URL https://dx.plos.org/10.1371/journal.pone.0086869.

[168] Robert Loddenkemper, Marc Lipman, and Alimuddin Zumla. Clinical aspects of adult tuberculosis. *Cold Spring Harbor Perspectives in Medicine*, 6 :a017848, 1 2016. ISSN 2157-1422. doi : 10.1101/cshperspect.a017848.

[169] Kerry H. Lok, William H. Benjamin, Michael E. Kimerling, Virginia Pruitt, Monica Lathan, Jafar Razeq, Nancy Hooper, Wendy Cronin, and Nancy E. Dunlap. Molecular Differentiation of Mycobacterium tuberculosis Strains without IS 6110 Insertions. *Emerging Infectious Diseases*, 8(11) :1310–1313, nov 2002. ISSN 1080-6040. doi : 10.3201/eid0811.020291. URL http://wwwnc.cdc.gov/eid/article/8/11/02-0291{_}article.htm.

[170] Simona Luca and Traian Mihaescu. History of bcg vaccine. *Maedica*, 8 :53–8, 3 2013. ISSN 1841-9038.

[171] Tao Luo, Iñaki Comas, Dan Luo, Bing Lu, Jie Wu, Lanhai Wei, Chongguang Yang, Qingyun Liu, Mingyu Gan, Gang Sun, Xin Shen, Feiying Liu, Sebastien Gagneux, Jian Mei, Rushu Lan, Kanglin Wan, and Qian Gao. Southern east asian origin and coexpansion of mycobacterium tuberculosis beijing family with han chinese. *Proc. Natl. Acad. Sci. U. S. A.*, 112(26) :8136–8141, June 2015.

[172] G G Mahairas, P J Sabo, M J Hickey, D C Singh, and C K Stover. Molecular analysis of genetic differences between Mycobacterium bovis BCG and virulent M. bovis. *Journal of bacteriology*, 178(5) :1274–1282, 1996. ISSN 0021-9193. doi : 10.1128/JB.178.5.1274-1282.1996. URL https://jb.asm.org/content/178/5/1274.

[173] Kira S. Makarova, Yuri I. Wolf, Omer S. Alkhnbashi, Fabrizio Costa, Shiraz A. Shah, Sita J. Saunders, Rodolphe Barrangou, Stan J. J. Brouns, Emmanuelle Charpentier, Daniel H. Haft, Philippe Horvath, Sylvain Moineau, Francisco J. M. Mojica, Rebecca M. Terns, Michael P. Terns, Malcolm F. White, Alexander F. Yakunin, Roger A. Garrett, John van der Oost, Rolf Backofen, and Eugene V. Koonin. An updated evolutionary classification of crispr–cas systems. *Nature Reviews Microbiology*, 13 :722–736, 11 2015. ISSN 1740-1526. doi : 10.1038/nrmicro3569.

[174] Magali Marmiesse, Priscille Brodin, Carmen Buchrieser, Christina Gutierrez, Nathalie Simoes, Veronique Vincent, Philippe Glaser, Stewart T. Cole, and Roland Brosch. Macro-array and bioinformatic analyses reveal mycobacterial 'core' genes, variation

in the ESAT-6 gene family and new phylogenetic markers for the Mycobacterium tuberculosis complex. *Microbiology*, 150(2) :483–496, feb 2004. ISSN 1350-0872. doi : 10.1099/mic.0.26662-0. URL https://www.microbiologyresearch.org/content/journal/micro/10.1099/mic.0.26662-0.

[175] Conor J. Meehan, Galo A. Goig, Thomas A. Kohl, Lennert Verboven, Anzaan Dippenaar, Matthew Ezewudo, Maha R. Farhat, Jennifer L. Guthrie, Kris Laukens, Paolo Miotto, Boatema Ofori-Anyinam, Viola Dreyer, Philip Supply, Anita Suresh, Christian Utpatel, Dick van Soolingen, Yang Zhou, Philip M. Ashton, Daniela Brites, Andrea M. Cabibbe, Bouke C. de Jong, Margaretha de Vos, Fabrizio Menardo, Sebastien Gagneux, Qian Gao, Tim H. Heupink, Qingyun Liu, Chloé Loiseau, Leen Rigouts, Timothy C. Rodwell, Elisa Tagliani, Timothy M. Walker, Robin M. Warren, Yanlin Zhao, Matteo Zignol, Marco Schito, Jennifer Gardy, Daniela M. Cirillo, Stefan Niemann, Inaki Comas, and Annelies Van Rie. Whole genome sequencing of Mycobacterium tuberculosis : current standards and open issues. *Nature Reviews Microbiology*, 17(9) :533–545, sep 2019. ISSN 1740-1526. doi : 10.1038/s41579-019-0214-5. URL http://www.nature.com/articles/s41579-019-0214-5.

[176] Kristin Kremer Mehrnoosh Doroudchi. IS6110-RFLP and Spoligotyping of Mycobacterium tuberculosis Isolates in Iran. *Scandinavian Journal of Infectious Diseases*, 32(6) :663–668, jan 2000. ISSN 0036-5548. doi : 10.1080/003655400459595. URL http://www.tandfonline.com/doi/full/10.1080/003655400459595.

[177] Fabrizio Menardo, Sebastian Duchêne, Daniela Brites, and Sebastien Gagneux. The molecular clock of mycobacterium tuberculosis. *PLoS Pathog.*, 15(9) :e1008067, September 2019.

[178] Fabrizio Menardo, Liliana K Rutaihwa, Michaela Zwyer, Sonia Borrell, Iñaki Comas, Emilyn Costa Conceição, Mireia Coscolla, Helen Cox, Moses Joloba, Horng-Yunn Dou, Julia Feldmann, Lukas Fenner, Janet Fyfe, Qian Gao, Darío García de Viedma, Alberto L Garcia-Basteiro, Sebastian M Gygli, Jerry Hella, Hellen Hiza, Levan Jugheli, Lujeko Kamwela, Midori Kato-Maeda, Qingyun Liu, Serej D Ley, Chloe Loiseau, Surakameth Mahasirimongkol, Bijaya Malla, Prasit Palittapongarnpim, Niaina Rakotosamimanana, Voahangy Rasolofo, Miriam Reinhard, Klaus Reither, Mohamed Sasamalo, Rafael Silva Duarte, Christophe Sola, Philip Suffys, Karla Valeria Batista Lima, Dorothy Yeboah-Manu, Christian Beisel, Daniela Brites, and Sebastien Gagneux. Local adaptation in populations of mycobacterium tuberculosis endemic to the indian ocean rim. *F1000Res.*, 10 :60, February 2021.

[179] Matthias Merker, Camille Blin, Stefano Mona, Nicolas Duforet-Frebourg, Sophie Lecher, Eve Willery, Michael G B Blum, Sabine Rüsch-Gerdes, Igor Mokrousov, Eman Aleksic, Caroline Allix-Béguec, Annick Antierens, Ewa Augustynowicz-Kopeć, Marie Ballif, Francesca Barletta, Hans Peter Beck, Clifton E Barry, 3rd, Maryline Bonnet, Emanuele Borroni, Isolina Campos-Herrero, Daniela Cirillo, Helen Cox, Suzanne Crowe, Valeriu Crudu, Roland Diel, Francis Drobniewski, Maryse Fauville-Dufaux, Sébastien Gagneux, Solomon Ghebremichael, Madeleine Hanekom, Sven Hoffner, Wei-Wei Jiao, Stobdan Kalon, Thomas A Kohl, Irina Kontsevaya, Troels Lillebæk, Shinji Maeda, Vladyslav Nikolayevskyy, Michael Rasmussen, Nalin Rastogi, Sofia Samper, Elisabeth Sanchez-Padilla, Branislava Savic, Isdore Chola Shamputa, Adong Shen, Li-Hwei Sng, Petras Stakenas, Kadri Toit, Francis Varaine, Dragana Vukovic, Céline Wahl, Robin Warren, Philip Supply, Stefan Niemann, and Thierry Wirth. Evolutionary history and global spread of the mycobacterium tuberculosis beijing lineage. *Nat. Genet.*, 47(3) :242–249, March 2015.

[180] Matthias Merker, Thomas A. Kohl, Stefan Niemann, and Philip Supply. *The Evo-*

lution of Strain Typing in the Mycobacterium tuberculosis Complex, pages 43–78. 2017. doi : 10.1007/978-3-319-64371-7_3.

[181] Matthias Merker, Maxime Barbier, Helen Cox, Jean-Philippe Rasigade, Silke Feuerriegel, Thomas Andreas Kohl, Roland Diel, Sonia Borrell, Sebastien Gagneux, Vladyslav Nikolayevskyy, Sönke Andres, Ulrich Nübel, Philip Supply, Thierry Wirth, and Stefan Niemann. Compensatory evolution drives multidrug-resistant tuberculosis in central asia. *Elife*, 7, October 2018.

[182] Olga Mestre, Tao Luo, Tiago Dos Vultos, Kristin Kremer, Alan Murray, Amine Namouchi, Céline Jackson, Jean Rauzier, Pablo Bifani, Rob Warren, Voahangy Rasolofo, Jian Mei, Qian Gao, and Brigitte Gicquel. Phylogeny of mycobacterium tuberculosis beijing strains constructed from polymorphisms in genes involved in DNA replication, recombination and repair. *PLoS One*, 6(1) :e16020, January 2011.

[183] Adrien Le Meur, Rima Zein-Eddine, Ombeline Lamer, Fiona Hak, Gaëtan Senelle, Jean-Philippe Vernadet, Samuel O'Donnell, Ricardo Rodriguez de la Vega, and Guislaine Refrégier. *Tools for short variant calling and the way to deal with big datasets*, pages 219–250. Elsevier, 2024. doi : 10.1016/B978-0-323-99886-4.00007-7.

[184] I. Milne, G. Stephen, M. Bayer, P. J. A. Cock, L. Pritchard, L. Cardle, P. D. Shaw, and D. Marshall. Using tablet for visual exploration of second-generation sequencing data. *Briefings in Bioinformatics*, 14 :193–202, 3 2013. ISSN 1467-5463. doi : 10.1093/bib/bbs012.

[185] Fuminori Mizukoshi, Tohru Miyoshi-Akiyama, Hiroki Iwai, Takako Suzuki, Reiko Kiritani, Teruo Kirikae, and Keiji Funatogawa. Genetic diversity of mycobacterium tuberculosis isolates from tochigi prefecture, a local region of japan. *BMC Infect. Dis.*, 17(1), December 2017.

[186] Francisco J. M. Mojica, Cesar Diez-Villasenor, Elena Soria, and Guadalupe Juez. Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Molecular Microbiology*, 36(1) :244–246, apr 2000. ISSN 0950-382X. doi : 10.1046/j.1365-2958.2000.01838.x. URL http://doi.wiley.com/10.1046/j.1365-2958.2000.01838.x.

[187] Igor Mokrousov. Genetic geography of mycobacterium tuberculosis beijing genotype : a multifacet mirror of human history ? *Infect. Genet. Evol.*, 8(6) :777–785, December 2008.

[188] Igor Mokrousov, Ho Minh Ly, Tatiana Otten, Nguyen Ngoc Lan, Boris Vyshnevskyi, Sven Hoffner, and Olga Narvskaya. Origin and primary dispersal of the mycobacterium tuberculosis beijing genotype : clues from human phylogeography. *Genome Res.*, 15(10) :1357–1364, October 2005.

[189] Igor Mokrousov, Olga Narvskaya, Anna Vyazovaya, Tatiana Otten, Wei-Wei Jiao, Lia Lima Gomes, Philip N Suffys, A-Dong Shen, and Boris Vishnevsky. Russian "successful" clone B0/W148 of mycobacterium tuberculosis beijing genotype : a multiplex PCR assay for rapid detection and global screening. *J. Clin. Microbiol.*, 50(11) :3757–3759, November 2012.

[190] Felix Mölder, Kim Philipp Jablonski, Brice Letcher, Michael B. Hall, Christopher H. Tomkins-Tinch, Vanessa Sochat, Jan Forster, Soohyun Lee, Sven O. Twardziok, Alexander Kanitz, Andreas Wilm, Manuel Holtgrewe, Sven Rahmann, Sven Nahnsen, and Johannes Köster. Sustainable data analysis with Snakemake. *F1000Research*, 10 :33, jan 2021. ISSN 2046-1402. doi : 10.12688/f1000research.29032.1. URL https://f1000research.com/articles/10-33/v1.

[191] Serge Mostowy, Debby Cousins, and Marcel A. Behr. Genomic interrogation of the dassie bacillus reveals it as a unique rd1 mutant within the mycobacterium tuberculosis complex. *Journal of Bacteriology*, 186 :104–109, 1 2004. ISSN 0021-9193. doi : 10.1128/JB.186.1.104-109.2003.

[192] Serge Mostowy, Anthony Onipede, Sebastien Gagneux, Stefan Niemann, Kristin Kremer, Edward P Desmond, Midori Kato-Maeda, and Marcel Behr. Genomic analysis distinguishes mycobacterium africanum. *Journal of clinical microbiology*, 42 :3594–9, 8 2004. ISSN 0095-1137. doi : 10.1128/JCM.42.8.3594-3599.2004.

[193] Serge Mostowy, Jackie Inwald, Steve Gordon, Carlos Martin, Rob Warren, Kristin Kremer, Debby Cousins, and Marcel A. Behr. Revisiting the evolution of mycobacterium bovis. *Journal of Bacteriology*, 187 :6386–6395, 9 2005. ISSN 0021-9193. doi : 10.1128/JB.187.18.6386-6395.2005.

[194] Claire V Mulholland, Abigail C Shockey, Htin L Aung, Ray T Cursons, Ronan F O'Toole, Sanjay S Gautam, Daniela Brites, Sebastien Gagneux, Sally A Roberts, Noel Karalus, Gregory M Cook, Caitlin S Pepperell, and Vickery L Arcus. Dispersal of mycobacterium tuberculosis driven by historical european trade in the south pacific. *Front. Microbiol.*, 10 :2778, December 2019.

[195] Borna Müller, Salome Dürr, Silvia Alonso, Jan Hattendorf, Cláudio J.M. Laisse, Sven D.C. Parsons, Paul D. van Helden, and Jakob Zinsstag. Zoonotic Mycobacterium bovis –induced Tuberculosis in Humans. *Emerging Infectious Diseases*, 19(6) :899–908, jun 2013. ISSN 1080-6040. doi : 10.3201/eid1906.120543. URL http://wwwnc.cdc.gov/eid/article/19/6/12-0543{_}article.htm.

[196] Borna Müller, Markus Hilty, Stefan Berg, M Carmen Garcia-Pelayo, James Dale, M Laura Boschiroli, Simeon Cadmus, Bongo Naré Richard Ngandolo, Sylvain Godreuil, Colette Diguimbaye-Djaibé, Rudovick Kazwala, Bassirou Bonfoh, Betty M Njanpop-Lafourcade, Naima Sahraoui, Djamel Guetarni, Abraham Aseffa, Meseret H Mekonnen, Voahangy Rasolofo Razanamparany, Herimanana Ramarokoto, Berit Djønne, James Oloya, Adelina Machado, Custodia Mucavele, Eystein Skjerve, Francoise Portaels, Leen Rigouts, Anita Michel, Annélle Müller, Gunilla Källenius, Paul D van Helden, R Glyn Hewinson, Jakob Zinsstag, Stephen V Gordon, and Noel H Smith. African 1, an epidemiologically important clonal complex of mycobacterium bovis dominant in mali, nigeria, cameroon, and chad. *Journal of bacteriology*, 191 :1951–60, 3 2009. ISSN 1098-5530. doi : 10.1128/JB.01590-08.

[197] Noriko Nakanishi, Takayuki Wada, Kentaro Arikawa, Julie Millet, Nalin Rastogi, and Tomotada Iwamoto. Evolutionary robust SNPs reveal the misclassification of mycobacterium tuberculosis beijing family strains into sublineages. *Infect. Genet. Evol.*, 16 :174–177, June 2013.

[198] A Nakata, M Amemura, and K Makino. Unusual nucleotide arrangement with repeated sequences in the Escherichia coli K-12 chromosome. *Journal of Bacteriology*, 171(6) :3553–3556, 1989. ISSN 0021-9193. doi : 10.1128/JB.171.6.3553-3556.1989. URL https://jb.asm.org/content/171/6/3553.

[199] Gary Napier, Susana Campino, Yared Merid, Markos Abebe, Yimtubezinash Woldeamanuel, Abraham Aseffa, Martin L Hibberd, Jody Phelan, and Taane G Clark. Robust barcoding and identification of mycobacterium tuberculosis lineages for epidemiological and clinical studies. *Genome Med.*, 12(1) :114, December 2020.

[200] Gary Napier, Susana Campino, Yared Merid, Markos Abebe, Yimtubezinash Woldeamanuel, Abraham Aseffa, Martin L Hibberd, Jody Phelan, and Taane G Clark. Robust barcoding and identification of mycobacterium tuberculosis lineages for epidemiological and clinical studies. *Genome medicine*, 12(1) :1–10, 2020.

[201] Hanna Nebenzahl-Guimaraes, Solomon A Yimer, Carol Holm-Hansen, Jessica de Beer, Roland Brosch, and Dick van Soolingen. Genomic characterization of Mycobacterium tuberculosis lineage 7 and a proposed name : 'Aethiops vetus'. *Microbial Genomics*, 2(6), jun 2016. ISSN 2057-5858. doi : 10.1099/mgen.0.000063. URL https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000063.

[202] Jean Claude Semuto Ngabonziza, Chloé Loiseau, Michael Marceau, Agathe Jouet, Fabrizio Menardo, Oren Tzfadia, Rudy Antoine, Esdras Belamo Niyigena, Wim Mulders, Kristina Fissette, Maren Diels, Cyril Gaudin, Stéphanie Duthoy, Willy Ssengooba, Emmanuel André, Michel K. Kaswa, Yves Mucyo Habimana, Daniela Brites, Dissou Affolabi, Jean Baptiste Mazarati, Bouke Catherine de Jong, Leen Rigouts, Sebastien Gagneux, Conor Joseph Meehan, and Philip Supply. A sister lineage of the Mycobacterium tuberculosis complex discovered in the African Great Lakes region. *Nature Communications*, 11(1) :2917, dec 2020. ISSN 2041-1723. doi : 10.1038/s41467-020-16626-6. URL http://www.nature.com/articles/s41467-020-16626-6.

[203] Jean Claude Semuto Ngabonziza, Chloé Loiseau, Michael Marceau, Agathe Jouet, Fabrizio Menardo, Oren Tzfadia, Rudy Antoine, Esdras Belamo Niyigena, Wim Mulders, Kristina Fissette, et al. A sister lineage of the mycobacterium tuberculosis complex discovered in the african great lakes region. *Nature communications*, 11 (1) :2917, 2020.

[204] World Health Organization. Issues relating to the use of bcg in immunization programmes, 1999.

[205] World Health Organization. Global tuberculosis report 2023, 2023. URL https://www.who.int/publications/i/item/9789240083851.

[206] M Ota, Y Hoshino, and S Hirao. Analysis of 605 tuberculosis outbreaks in japan, 1993-2015 : time, place and transmission site. *Epidemiol. Infect.*, 149(e85) :e85, March 2021.

[207] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. 3 2022.

[208] Louise J Pankhurst, Carlos del Ojo Elias, Antonina A Votintseva, Timothy M Walker, Kevin Cole, Jim Davies, Jilles M Fermont, Deborah M Gascoyne-Binzi, Thomas A Kohl, Clare Kong, Nadine Lemaitre, Stefan Niemann, John Paul, Thomas R Rogers, Emma Roycroft, E Grace Smith, Philip Supply, Patrick Tang, Mark H Wilcox, Sarah Wordsworth, David Wyllie, Li Xu, and Derrick W Crook. Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing : a prospective study. *The Lancet Respiratory Medicine*, 4 :49–58, 1 2016. ISSN 22132600. doi : 10.1016/S2213-2600(15)00466-X.

[209] Young Kil Park, Heeyoon Kang, Heekyung Yoo, Seung Heon Lee, Hanseong Roh, Hee Jin Kim, and Sungweon Ryoo. Whole-Genome sequence of mycobacterium tuberculosis korean strain KIT87190. *Genome Announc.*, 2(5), October 2014.

[210] Sven D.C. Parsons, Julian A. Drewe, Nicolaas C. Gey van Pittius, Robin M. Warren, and Paul D. van Helden. N ovel cause of tuberculosis in meerkats, south africa. *Emerging Infectious Diseases*, 19 :2004–2007, 12 2013. ISSN 1080-6040. doi : 10.3201/eid1912.130268.

[211] Xiuli Peng and Jianjun Sun. Mechanism of ESAT-6 membrane interaction and its roles in pathogenesis of Mycobacterium tuberculosis. *Toxicon*, 116 :29–34, jun 2016. ISSN 00410101. doi : 10.1016/j.toxicon.2015.10.003. URL https://linkinghub.elsevier.com/retrieve/pii/S0041010115301069.

[212] Caitlin S Pepperell, Julie M Granka, David C Alexander, Marcel A Behr, Linda Chui, Janet Gordon, Jennifer L Guthrie, Frances B Jamieson, Deanne Langlois-Klassen, Richard Long, Dao Nguyen, Wendy Wobeser, and Marcus W Feldman. Dispersal of mycobacterium tuberculosis via the canadian fur trade. *Proc. Natl. Acad. Sci. U. S. A.*, 108(16) :6526–6531, April 2011.

[213] Jody E Phelan, Denise M O'Sullivan, Diana Machado, Jorge Ramos, Yaa EA Oppong, Susana Campino, Justin O'Grady, Ruth McNerney, Martin L Hibberd, Miguel Viveiros, et al. Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. *Genome medicine*, 11(1) : 1–7, 2019.

[214] Andrey Prjibelski, Dmitry Antipov, Dmitry Meleshko, Alla Lapidus, and Anton Korobeynikov. Using spades de novo assembler. *Current Protocols in Bioinformatics*, 70, 6 2020. ISSN 1934-3396. doi : 10.1002/cpbi.102.

[215] Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. Rocketqa : An optimized training approach to dense passage retrieval for open-domain question answering. 10 2020.

[216] Aaron R. Quinlan and Ira M. Hall. BEDTools : a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6) :841–842, mar 2010. ISSN 1460-2059. doi : 10.1093/bioinformatics/btq033. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq033.

[217] Marcelo Fouad Rabahi, José Laerte Rodrigues da Silva Júnior, Anna Carolina Galvão Ferreira, Daniela Graner Schuwartz Tannus-Silva, and Marcus Barreto Conde. Tuberculosis treatment. *Jornal brasileiro de pneumologia : publicacao oficial da Sociedade Brasileira de Pneumologia e Tisilogia*, 43 :472–486, 2017. ISSN 1806-3756. doi : 10.1590/S1806-37562016000000388.

[218] Mina Ebrahimi Rad, Pablo Bifani, Carlos Martin, Kristin Kremer, Sofia Samper, Jean Rauzier, Barry Kreiswirth, Jesus Blazquez, Marc Jouan, Dick Van Soolingen, et al. Mutations in putative mutator genes of mycobacterium tuberculosis strains of the w-beijing family. *Emerging infectious diseases*, 9(7) :838, 2003.

[219] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018. URL https://api.semanticscholar.org/CorpusID:49313245.

[220] Rahim Rajwani, Sheeba Shehzad, and Gilman Kit Hang Siu. Miru-profiler : a rapid tool for determination of 24-loci miru-vntr profiles from assembled genomes of mycobacterium tuberculosis. *PeerJ*, 6 :e5090, 2018.

[221] Jean Rauzier, Eamonn Gormley, M. Cristina Gutierrez, Eric Kassa-Kelembho, Laurie J. Sandall, Chris Dupont, Brigitte Gicquel, and Alan Murray. A novel polymorphic genetic locus in members of the mycobacterium tuberculosis complex. *Microbiology*, 145 :1695–1701, 7 1999. ISSN 1350-0872. doi : 10.1099/13500872-145-7-1695.

[222] Niraj Rayamajhi, Chi-Hing Christina Cheng, and Julian M Catchen. Evaluating Illumina-, Nanopore-, and PacBio-based genome assembly strategies with the bald notothen, Trematomus borchgrevinki. *G3 Genes|Genomes|Genetics*, 12(11) :jkac192,

07 2022. ISSN 2160-1836. doi : 10.1093/g3journal/jkac192. URL https://doi.org/10.1093/g3journal/jkac192.

[223] T. B. K. Reddy, R. Riley, F. Wymore, P. Montgomery, D. DeCaprio, R. Engels, M. Gellesch, J. Hubble, D. Jen, H. Jin, M. Koehrsen, L. Larson, M. Mao, M. Nitzberg, P. Sisk, C. Stolte, B. Weiner, J. White, Z. K. Zachariah, G. Sherlock, J. E. Galagan, C. A. Ball, and G. K. Schoolnik. Tb database : an integrated platform for tuberculosis research. *Nucleic Acids Research*, 37 :D499–D508, 1 2009. ISSN 0305-1048. doi : 10.1093/nar/gkn652.

[224] Guislaine Refrégier, Edgar Abadia, Tomoshige Matsumoto, Hiromi Ano, Tetsuya Takashima, Izuo Tsuyuguchi, Elif Aktas, Füsun Cömert, Michel Kireopori Gomgnimbou, Stefan Panaiotov, Jody Phelan, Francesc Coll, Ruth McNerney, Arnab Pain, Taane G Clark, and Christophe Sola. Turkish and japanese mycobacterium tuberculosis sublineages share a remote common ancestor. *Infect. Genet. Evol.*, 45 : 461–473, November 2016.

[225] Guislaine Refrégier, Christophe Sola, and Christophe Guyeux. Unexpected diversity of CRISPR unveils some evolutionary patterns of repeated sequences in mycobacterium tuberculosis. *BMC Genomics*, 21(1) :841, November 2020.

[226] J T Rhee, M M Tanaka, M A Behr, C B Agasino, E A Paz, P C Hopewell, and P M Small. Use of multiple markers in population-based molecular epidemiologic studies of tuberculosis. *The International Journal of Tuberculosis and Lung Disease*, 4(12) : 1111–1119, 2000.

[227] Stefan Riedel, Stephen A Morse, Timothy A Mietzner, and Steve Miller. *Jawetz, Melnick & Adelberg's medical microbiology*. 2019. ISBN 1260012026 9781260012026.

[228] Huanwei Ru, Xiaojia Liu, Chen Lin, Jingyan Yang, Fuzeng Chen, Ruifeng Sun, Lu Zhang, and Jun Liu. The Impact of Genome Region of Difference 4 (RD4) on Mycobacterial Virulence and BCG Efficacy. *Frontiers in Cellular and Infection Microbiology*, 7, jun 2017. ISSN 2235-2988. doi : 10.3389/fcimb.2017.00239. URL http://journal.frontiersin.org/article/10.3389/fcimb.2017.00239/full.

[229] Kenneth J Ryan. Sherris Medical Microbiology, 7e, 2018. URL https://login.proxy.bib.uottawa.ca/login?url=https://accessmedicine.mhmedical.com/book.aspx?bookid=2268.

[230] Muhammed Rabiu Sahal, Gaetan Senelle, Kevin La, Barbara Molina-Moya, Jose Dominguez, Tukur Panda, Emmanuelle Cambau, Guislaine Refregier, Christophe Sola, and Christophe Guyeux. An updated evolutionary history and taxonomy of mycobacterium tuberculosis lineage 5, also called m. africanum. *bioRxiv*, 2023. doi : 10.1101/2022.11.21.517336. URL https://www.biorxiv.org/content/early/2023/02/24/2022.11.21.517336.

[231] Muhammed Rabiu Sahal, Gaetan Senelle, Kevin La, Tukur Wada Panda, Dalha Wada Taura, Christophe Guyeux, Emmanuelle Cambau, and Christophe Sola. Mycobacterium tuberculosis complex drug-resistance, phylogenetics, and evolution in nigeria : Comparison with ghana and cameroon. *PLOS Neglected Tropical Diseases*, 17 :e0011619, 10 2023. ISSN 1935-2735. doi : 10.1371/journal.pntd.0011619.

[232] H. Salamon. Detection of deleted genomic dna using a semiautomated computational analysis of genechip data. *Genome Research*, 10 :2044–2054, 12 2000. ISSN 10889051. doi : 10.1101/gr.GR-1529R.

[233] Anita C Schürch, Kristin Kremer, Amber C A Hendriks, Benthe Freyee, Christopher R E McEvoy, Reinout van Crevel, Martin J Boeree, Paul van Helden, Robin M Warren, Roland J Siezen, and Dick van Soolingen. Snp/rd typing of mycobacterium tuberculosis beijing strains reveals local and worldwide disseminated clonal complexes. *PloS one*, 6 :e28365, 2011. ISSN 1932-6203. doi : 10.1371/journal.pone.0028365.

[234] Torsten Seemann. shovill. URL https://github.com/tseemann/shovill.

[235] Torsten Seemann. snippy : fast bacterial variant calling from NGS reads, 2015. URL https://github.com/tseemann/snippy.

[236] Torsten Seemann. samclip, 2018. URL https://github.com/tseemann/samclip.

[237] Gaetan Senelle, Christophe Guyeux, Guislaine Refrégier, and Christophe Sola. Investigating the diversity of tuberculosis spoligotypes with dimensionality reduction and graph theory. *Genes*, 13, 12 2022. ISSN 2073-4425. doi : 10.3390/genes13122328.

[238] Gaetan Senelle, Muhammed Rabiu Sahal, Kevin La, Barbara Molina-Moya, Jose Dominguez, Tukur Panda, Emmanuelle Cambau, Guislaine Refregier, Christophe Sola, and Christophe Guyeux. An updated evolutionary history and taxonomy of mycobacterium tuberculosis lineage 5, also called m. africanum. *bioRxiv*, 2022. doi : 10.1101/2022.11.21.517336. URL https://www.biorxiv.org/content/early/2022/11/21/2022.11.21.517336.

[239] Gaëtan Senelle, Christophe Guyeux, Emmanuelle Cambau, Guislaine Refrégier, and Christophe Sola. Tb-annotator : A scalable web application that allows in-depth analysis of very large sets of publicly available mycobacterium tuberculosis complex genomes. In *43rd Annual Congress of the European Society of Mycobacteriology (ESM 2023)*, Tirana, Albania, jun 2023.

[240] Gaetan Senelle, Christophe Guyeux, Guislaine Refrégier, and Christophe Sola. Tb-annotator : a scalable web application that allows in-depth analysis of very large sets of publicly available mycobacterium tuberculosis complex genomes. *bioRxiv*, 2023. doi : 10.1101/2023.06.12.526393. URL https://www.biorxiv.org/content/early/2023/06/13/2023.06.12.526393.

[241] Gaetan Senelle, Christophe Guyeux, Guislaine Refrégier, and Christophe Sola. Towards a better understanding of the long-lasting evolutionary history of mycobacterium tuberculosis. *Tuberculosis*, 143 :102374, 12 2023. ISSN 14729792. doi : 10.1016/j.tube.2023.102374.

[242] Surendra K Sharma and Alladi Mohan. Miliary tuberculosis. *Microbiology spectrum*, 5, 3 2017. ISSN 2165-0497. doi : 10.1128/microbiolspec.TNMI7-0013-2016.

[243] Wei Shen, Shuai Le, Yan Li, and Fuquan Hu. Seqkit : A cross-platform and ultrafast toolkit for fasta/q file manipulation. *PLOS ONE*, 11 :e0163962, 10 2016. ISSN 1932-6203. doi : 10.1371/journal.pone.0163962.

[244] Egor Shitikov, Sergey Kolchenko, Igor Mokrousov, Julia Bespyatykh, Dmitry Ischenko, Elena Ilina, and Vadim Govorun. Evolutionary pathway analysis and unified classification of east asian lineage of mycobacterium tuberculosis. *Scientific Reports*, 7 :9227, 8 2017. ISSN 2045-2322. doi : 10.1038/s41598-017-10018-5.

[245] Egor Shitikov, Sergey Kolchenko, Igor Mokrousov, Julia Bespyatykh, Dmitry Ischenko, Elena Ilina, and Vadim Govorun. Evolutionary pathway analysis and unified classification of east asian lineage of mycobacterium tuberculosis. *Sci. Rep.*, 7(1) :9227, August 2017.

[246] Egor Shitikov, Andrei Guliaev, Julia Bespyatykh, Maja Malakhova, Sergey Kolchenko, Georgy Smirnov, Matthias Merker, Stefan Niemann, Igor Mokrousov, Elena Ilina, and Vadim Govorun. The role of IS6110 in micro- and macroevolution of mycobacterium tuberculosis lineage 2. *Mol. Phylogenet. Evol.*, 139(106559) :106559, October 2019.

[247] Egor Shitikov, Andrei Guliaev, Julia Bespyatykh, Maja Malakhova, Sergey Kolchenko, Georgy Smirnov, Matthias Merker, Stefan Niemann, Igor Mokrousov, Elena Ilina, et al. The role of is6110 in micro-and macroevolution of mycobacterium tuberculosis lineage 2. *Molecular phylogenetics and evolution*, 139 :106559, 2019.

[248] G. B. Smirnov. Repeats in bacterial genome : Evolutionary considerations. *Molecular Genetics, Microbiology and Virology*, 25 :56–65, 6 2010. ISSN 0891-4168. doi : 10.3103/S0891416810020023.

[249] Noel H. Smith, Kristin Kremer, Jacqueline Inwald, James Dale, Jeffrey R. Driscoll, Stephen V. Gordon, Dick van Soolingen, R. Glyn Hewinson, and John Maynard Smith. Ecotypes of the Mycobacterium tuberculosis complex. *Journal of Theoretical Biology*, 239(2) :220–225, mar 2006. ISSN 00225193. doi : 10.1016/j.jtbi.2005.08.036. URL https://linkinghub.elsevier.com/retrieve/pii/S0022519305003838.

[250] Noel H. Smith, Stefan Berg, James Dale, Adrian Allen, Sabrina Rodriguez, Beatriz Romero, Filipa Matos, Solomon Ghebremichael, Claudine Karoui, Chiara Donati, Adelina da Conceicao Machado, Custodia Mucavele, Rudovick R. Kazwala, Markus Hilty, Simeon Cadmus, Bongo Naré Richard Ngandolo, Meseret Habtamu, James Oloya, Annélle Muller, Feliciano Milian-Suazo, Olga Andrievskaia, Michaela Projahn, Soledad Barandiarán, Analía Macías, Borna Müller, Marcos Santos Zanini, Cassia Yumi Ikuta, Cesar Alejandro Rosales Rodriguez, Sônia Regina Pinheiro, Alvaro Figueroa, Sang-Nae Cho, Nader Mosavari, Pei-Chun Chuang, Ruwen Jou, Jakob Zinsstag, Dick van Soolingen, Eamonn Costello, Abraham Aseffa, Freddy Proaño-Perez, Françoise Portaels, Leen Rigouts, Angel Adrián Cataldi, Desmond M. Collins, María Laura Boschiroli, R. Glyn Hewinson, José Soares Ferreira Neto, Om Surujballi, Keyvan Tadyon, Ana Botelho, Ana María Zárraga, Nicky Buller, Robin Skuce, Anita Michel, Alicia Aranaz, Stephen V. Gordon, Bo-Young Jeon, Gunilla Källenius, Stefan Niemann, M. Beatrice Boniotti, Paul D. van Helden, Beth Harris, Martín José Zumárraga, and Kristin Kremer. European 1 : A globally important clonal complex of mycobacterium bovis. *Infection, Genetics and Evolution*, 11 :1340–1351, 8 2011. ISSN 15671348. doi : 10.1016/j.meegid.2011.04.027.

[251] Christophe Sola, Anne Devallois, Lionel Horgen, Jérôme Maïsetti, Ingrid Filliol, Eric Legrand, and Nalin Rastogi. Tuberculosis in the Caribbean : Using Spacer Oligonucleotide Typing to Understand Strain Origin and Transmission. *Emerging Infectious Diseases*, 5(3) :404–411, jun 1999. ISSN 1080-6040. doi : 10.3201/eid0503.990311. URL https://wwwnc.cdc.gov/eid/article/5/3/99-0311{_}article.

[252] Christophe Sola, Séverine Ferdinand, Leonardo A Sechi, Stefania Zanetti, Dominique Martial, Caterina Mammina, Antonino Nastasi, Giovanni Fadda, and Nalin Rastogi. Mycobacterium tuberculosis molecular evolution in western mediterranean islands of sicily and sardinia. *Infect. Genet. Evol.*, 5(2) :145–156, March 2005.

[253] Christophe Sola, Christophe Guyeux, and Gaëtan Senelle. Le chaos croissant de la génomique des populations des bacilles de la tuberculose à l'ère des big data. In *SFM Mycodays (2022)*, Lyon, France, jun 2022.

[254] Christophe Sola, Gaëtan Senelle, K. La, T. Billard-Pomares, J. Marin, A. Bridier-Nahmias, Christophe Guyeux, Guislaine Refrégier, E. Carbonnelle, and Emmanuelle Cambau. The growing chaos of tuberculosis population genomics at the era of 'big

data' : sorting out the wheat from the chaff. In *42nd Annual Congress of the European Society of Mycobacteriology (ESM 2022)*, Bologna, Italy, jun 2022. URL https://publiweb.femto-st.fr/tntnet/entries/19127/documents/author/data.

[255] Christophe Sola, Igor Mokrousov, Muhammed Rabiu Sahal, Kevin La, Gaetan Senelle, Christophe Guyeux, Guislaine Refrégier, and Emmanuelle Cambau. *Evolution, Phylogenetics, and Phylogeography of Mycobacterium tuberculosis complex*, pages 683–772. Elsevier, 2024. doi : 10.1016/B978-0-443-28818-0.00025-2.

[256] Rotem Sorek, Victor Kunin, and Philip Hugenholtz. Crispr — a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nature Reviews Microbiology*, 6 :181–186, 3 2008. ISSN 1740-1526. doi : 10.1038/nrmicro1793.

[257] Srinand Sreevatsan, Xi Pan, Kathryn E. Stockbauer, Nancy D. Connell, Barry N. Kreiswirth, Thomas S. Whittam, and James M. Musser. Restricted structural gene polymorphism in the mycobacterium tuberculosis complex indicates evolutionarily recent global dissemination. *Proceedings of the National Academy of Sciences*, 94 : 9869–9874, 9 1997. ISSN 0027-8424. doi : 10.1073/pnas.94.18.9869.

[258] Prapaporn Srilohasin, Therdsak Prammananan, Kiatichai Faksri, Jody E Phelan, Prapat Suriyaphol, Phalin Kamolwat, Saijai Smithtikarn, Areeya Disratthakit, Sanjib Mani Regmi, Manoon Leechawengwongs, Rick Twee-Hee Ong, Yik Ying Teo, Sissades Tongsima, Taane G Clark, and Angkana Chaiprasert. Genomic evidence supporting the clonal expansion of extensively drug-resistant tuberculosis bacteria belonging to a rare proto-beijing genotype. *Emerg. Microbes Infect.*, 9(1) :2632–2641, December 2020.

[259] Alexandros Stamatakis. Raxml version 8 : a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30 :1312–1313, 5 2014. ISSN 1367-4811. doi : 10.1093/bioinformatics/btu033.

[260] Alexandros Stamatakis. RAxML version 8 : a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9) :1312–1313, May 2014.

[261] W. Steenken. Lysis of tubercle bacilli in vitro. *Experimental Biology and Medicine*, 33 :253–255, 11 1935. ISSN 1535-3702. doi : 10.3181/00379727-33-8330P.

[262] Andreas Steiner, David Stucki, Mireia Coscolla, Sonia Borrell, and Sebastien Gagneux. Kvarq : targeted and direct variant calling from fastq reads of bacterial genomes. *BMC Genomics*, 15 :881, 12 2014. ISSN 1471-2164. doi : 10.1186/1471-2164-15-881.

[263] David Stucki and Sebastien Gagneux. Single nucleotide polymorphisms in Mycobacterium tuberculosis and the need for a curated database. *Tuberculosis*, 93 (1) :30–39, jan 2013. ISSN 14729792. doi : 10.1016/j.tube.2012.11.002. URL https://linkinghub.elsevier.com/retrieve/pii/S1472979212002028.

[264] David Stucki, Bijaya Malla, Simon Hostettler, Thembela Huna, Julia Feldmann, Dorothy Yeboah-Manu, Sonia Borrell, Lukas Fenner, Iñaki Comas, Mireia Coscollà, and Sebastien Gagneux. Two new rapid snp-typing methods for classifying mycobacterium tuberculosis complex into the main phylogenetic lineages. *PLoS ONE*, 7 : e41253, 7 2012. ISSN 1932-6203. doi : 10.1371/journal.pone.0041253.

[265] David Stucki, Marie Ballif, Thomas Bodmer, Mireia Coscolla, Anne-Marie Maurer, Sara Droz, Christa Butz, Sonia Borrell, Christel Längle, Julia Feldmann, Hansjakob Furrer, Carlo Mordasini, Peter Helbling, Hans L. Rieder, Matthias Egger, Sébastien Gagneux, and Lukas Fenner. Tracking a tuberculosis outbreak over 21

years : Strain-specific single-nucleotide polymorphism typing combined with targeted whole-genome sequencing. *The Journal of Infectious Diseases*, 211 :1306–1316, 4 2015. ISSN 1537-6613. doi : 10.1093/infdis/jiu601.

[266] David Stucki, Daniela Brites, Leïla Jeljeli, Mireia Coscolla, Qingyun Liu, Andrej Trauner, Lukas Fenner, Liliana Rutaihwa, Sonia Borrell, Tao Luo, et al. Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages. *Nature genetics*, 48(12) :1535–1543, 2016.

[267] Weiwei Sun, Zheng Chen, Xinyu Ma, Lingyong Yan, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. Instruction distillation makes large language models efficient zero-shot rankers. 11 2023.

[268] Philip Supply, Juana Magdalena, Sabine Himpens, and Camille Locht. Identification of novel intergenic repetitive units in a mycobacterial two-component system operon. *Molecular Microbiology*, 26 :991–1003, 12 1997. ISSN 0950-382X. doi : 10.1046/j.1365-2958.1997.6361999.x.

[269] Philip Supply, Sarah Lesjean, Evgueni Savine, Kristin Kremer, Dick Van Soolingen, and Camille Locht. Automated high-throughput genotyping for study of global epidemiology of mycobacterium tuberculosis based on mycobacterial interspersed repetitive units. *Journal of clinical microbiology*, 39(10) :3563–3571, 2001.

[270] Philip Supply, Caroline Allix, Sarah Lesjean, Mara Cardoso-Oelemann, Sabine Rüsch-Gerdes, Eve Willery, Evgueni Savine, Petra de Haas, Henk van Deutekom, Solvig Roring, Pablo Bifani, Natalia Kurepina, Barry Kreiswirth, Christophe Sola, Nalin Rastogi, Vincent Vatin, Maria Cristina Gutierrez, Maryse Fauville, Stefan Niemann, Robin Skuce, Kristin Kremer, Camille Locht, and Dick van Soolingen. Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of mycobacterium tuberculosis. *J. Clin. Microbiol.*, 44(12) :4498–4510, December 2006.

[271] Philip Supply, Caroline Allix, Sarah Lesjean, Mara Cardoso-Oelemann, Sabine Rüsch-Gerdes, Eve Willery, Evgueni Savine, Petra de Haas, Henk van Deutekom, Solvig Roring, Pablo Bifani, Natalia Kurepina, Barry Kreiswirth, Christophe Sola, Nalin Rastogi, Vincent Vatin, Maria Cristina Gutierrez, Maryse Fauville, Stefan Niemann, Robin Skuce, Kristin Kremer, Camille Locht, and Dick van Soolingen. Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of mycobacterium tuberculosis. *Journal of Clinical Microbiology*, 44 :4498–4510, 12 2006. ISSN 0095-1137. doi : 10.1128/JCM.01392-06.

[272] Shin Suzuki, Tomohiro Yasuda, Yuichi Shiraishi, Satoru Miyano, and Masao Nagasaki. ClipCrop : a tool for detecting structural variations with single-base resolution using soft-clipping information. *BMC Bioinformatics*, 12(Suppl 14) :S7, 2011. ISSN 1471-2105. doi : 10.1186/1471-2105-12-S14-S7. URL http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-S14-S7.

[273] T Suzuki and T Inoue. Earliest evidence of spinal tuberculosis from the aneolithic yayoi period in japan. *Int. J. Osteoarchaeol.*, 17(4) :392–402, July 2007.

[274] Takao Suzuki, Hisashi Fujita, and Jong Gyu Choi. Brief communication : new evidence of tuberculosis from prehistoric Korea-Population movement and early evidence of tuberculosis in far east asia. *Am. J. Phys. Anthropol.*, 136(3) :357–360, July 2008.

[275] Cheng Yee Tang and Rick Twee-Hee Ong. Mirureader : Miru-vntr typing directly from long sequencing reads. *Bioinformatics*, 36(5) :1625–1626, 2020.

[276] WHO Team. *Latent tuberculosis infection : updated and consolidated guidelines for programmatic management*. 2018.

[277] Yuttapong Thawornwattana, Surakameth Mahasirimongkol, Hideki Yanai, Htet Myat Win Maung, Zhezhe Cui, Virasakdi Chongsuvivatwong, and Prasit Palittapongarnpim. Revised nomenclature and SNP barcode for mycobacterium tuberculosis lineage 2. *Microb. Genom.*, 7(11), November 2021.

[278] D Thierry, A Brisson-Noël, V Vincent-Lévy-Frébault, S Nguyen, J L Guesdon, and B Gicquel. Characterization of a Mycobacterium tuberculosis insertion sequence, IS6110, and its application in diagnosis. *Journal of Clinical Microbiology*, 28(12) : 2668–2673, 1990. ISSN 0095-1137. doi : 10.1128/JCM.28.12.2668-2673.1990. URL https://jcm.asm.org/content/28/12/2668.

[279] Bernard Thomann. L'hygiène nationale, la société civile et la reconnaissance de la silicose comme maladie professionnelle au japon (1868-1960). *Rev. Hist. Mod. Contemp.*, 56-1(1) :142, 2009.

[280] Thomas M. Daniel. *Pioneers of Medicine and Their Impact on Tuberculosis*. 2000.

[281] A. G. Tsolaki, A. E. Hirsh, K. DeRiemer, J. A. Enciso, M. Z. Wong, M. Hannan, Y.-O. L. G. de la Salmoniere, K. Aman, M. Kato-Maeda, and P. M. Small. Functional and evolutionary genomics of Mycobacterium tuberculosis : Insights from genomic deletions in 100 strains. *Proceedings of the National Academy of Sciences*, 101(14) : 4865–4870, apr 2004. ISSN 0027-8424. doi : 10.1073/pnas.0305634101. URL http://www.pnas.org/cgi/doi/10.1073/pnas.0305634101.

[282] A. G. Tsolaki, S. Gagneux, A. S. Pym, Y.-O. L. Goguet de la Salmoniere, B. N. Kreiswirth, D. Van Soolingen, and P. M. Small. Genomic Deletions Classify the Beijing/W Strains as a Distinct Genetic Lineage of Mycobacterium tuberculosis. *Journal of Clinical Microbiology*, 43(7) :3185–3191, jul 2005. ISSN 0095-1137. doi : 10.1128/JCM.43.7.3185-3191.2005. URL https://jcm.asm.org/content/43/7/3185.

[283] Anthony G Tsolaki, Sebastien Gagneux, Alexander S Pym, Yves-Olivier L Goguet de la Salmoniere, Barry N Kreiswirth, Dick Van Soolingen, and Peter M Small. Genomic deletions classify the Beijing/W strains as a distinct genetic lineage of mycobacterium tuberculosis. *J. Clin. Microbiol.*, 43(7) :3185–3191, July 2005.

[284] J D van Embden, M D Cave, J T Crawford, J W Dale, K D Eisenach, B Gicquel, P Hermans, C Martin, R McAdam, and T M Shinnick. Strain identification of Mycobacterium tuberculosis by DNA fingerprinting : recommendations for a standardized methodology. *Journal of Clinical Microbiology*, 31(2) :406–409, 1993. ISSN 0095-1137. doi : 10.1128/JCM.31.2.406-409.1993. URL https://jcm.asm.org/content/31/2/406.

[285] J D van Embden, M D Cave, J T Crawford, J W Dale, K D Eisenach, B Gicquel, P Hermans, C Martin, R McAdam, and T M Shinnick. Strain identification of mycobacterium tuberculosis by DNA fingerprinting : recommendations for a standardized methodology. *J. Clin. Microbiol.*, 31(2) :406–409, February 1993.

[286] Jakko van Ingen, Zeaur Rahim, Arnout Mulder, Martin J. Boeree, Roxane Simeone, Roland Brosch, and Dick van Soolingen. Characterization of Mycobacterium orygis as M. tuberculosis Complex Subspecies. *Emerging Infectious Diseases*, 18(4) :653–655, apr 2012. ISSN 1080-6040. doi : 10.3201/eid1804.110888. URL http://wwwnc.cdc.gov/eid/article/18/4/11-0888{_}article.htm.

[287] D van Soolingen, P W Hermans, P E de Haas, D R Soll, and J D van Embden. Occurrence and stability of insertion sequences in Mycobacterium tuberculosis complex strains : evaluation of an insertion sequence-dependent DNA polymorphism as a tool in the epidemiology of tuberculosis. *Journal of Clinical Microbiology*, 29(11) : 2578–2586, 1991. ISSN 0095-1137. doi : 10.1128/JCM.29.11.2578-2586.1991. URL https://jcm.asm.org/content/29/11/2578.

[288] D van Soolingen, P E de Haas, P W Hermans, P M Groenen, and J D van Embden. Comparison of various repetitive DNA elements as genetic markers for strain differentiation and epidemiology of Mycobacterium tuberculosis. *Journal of Clinical Microbiology*, 31(8) :1987–1995, 1993. ISSN 0095-1137. doi : 10.1128/JCM.31.8.1987-1995.1993. URL https://jcm.asm.org/content/31/8/1987.

[289] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. 6 2017.

[290] C T Vidya and K P Prabheesh. Implications of COVID-19 pandemic on the global trade networks. *Emerg. Mark. Fin. Trade*, 56(10) :2408–2421, August 2020.

[291] Takayuki Wada and Tomotada Iwamoto. Allelic diversity of variable number of tandem repeats provides phylogenetic clues regarding the mycobacterium tuberculosis beijing family. *Infect. Genet. Evol.*, 9(5) :921–926, September 2009.

[292] Takayuki Wada, Tomotada Iwamoto, Atsushi Hase, and Shinji Maeda. Scanning of genetic diversity of evolutionarily sequential mycobacterium tuberculosis beijing family strains based on genome wide analysis. *Infect. Genet. Evol.*, 12(7) :1392–1396, October 2012.

[293] Takayuki Wada, Tomotada Iwamoto, Aki Tamaru, Junji Seto, Tadayuki Ahiko, Kaori Yamamoto, Atushi Hase, Shinji Maeda, and Taro Yamamoto. Clonality and micro-diversity of a nationwide spreading genotype of mycobacterium tuberculosis in japan. *PLoS One*, 10(3) :e0118495, March 2015.

[294] Timothy M Walker, Camilla LC Ip, Ruth H Harrell, Jason T Evans, Georgia Kapatai, Martin J Dedicoat, David W Eyre, Daniel J Wilson, Peter M Hawkey, Derrick W Crook, Julian Parkhill, David Harris, A Sarah Walker, Rory Bowden, Philip Monk, E Grace Smith, and Tim EA Peto. Whole-genome sequencing to delineate mycobacterium tuberculosis outbreaks : a retrospective observational study. *The Lancet Infectious Diseases*, 13 :137–146, 2 2013. ISSN 14733099. doi : 10.1016/S1473-3099(12)70277-3.

[295] Kanglin Wan, Jinghua Liu, Yolande Hauck, Yuanyuan Zhang, Jie Liu, Xiuqin Zhao, Zhiguang Liu, Bing Lu, Haiyan Dong, Yi Jiang, Kristin Kremer, Gilles Vergnaud, Dick van Soolingen, and Christine Pourcel. Investigation on mycobacterium tuberculosis diversity in china and the origin of the beijing clade. *PLoS One*, 6(12) : e29190, December 2011.

[296] J. D. WATSON and F. H. C. CRICK. Molecular Structure of Nucleic Acids : A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356) :737–738, apr 1953. ISSN 0028-0836. doi : 10.1038/171737a0. URL http://www.nature.com/articles/171737a0.

[297] Alice R Wattam, James J Davis, Rida Assaf, Sébastien Boisvert, Thomas Brettin, Christopher Bun, Neal Conrad, Emily M Dietrich, Terry Disz, Joseph L Gabbard, et al. Improvements to patric, the all-bacterial bioinformatics database and analysis resource center. *Nucleic acids research*, 45(D1) :D535–D542, 2017.

[298] HEINZ ULRICH WEIER and JOE W. GRAY. A Programmable System to Perform the Polymerase Chain Reaction. *DNA*, 7(6) :441–447, jul 1988. ISSN 0198-0238. doi : 10.1089/dna.1.1988.7.441. URL http://www.liebertpub.com/doi/10.1089/dna.1.1988.7.441.

[299] T. Weniger, J. Krawczyk, P. Supply, S. Niemann, and D. Harmsen. Miru-vntrplus : a web tool for polyphasic genotyping of mycobacterium tuberculosis complex bacteria. *Nucleic Acids Research*, 38 :W326–W331, 7 2010. ISSN 0305-1048. doi : 10.1093/nar/gkq351.

[300] WHO. Bcg vaccines : Who position paper – february 2018, 2018.

[301] WHO. Meeting report of the who expert consultation on the definition of extensively drug-resistant tuberculosis, 2020.

[302] WHO. Catalogue of mutations in mycobacterium tuberculosis complex and their association with drug resistance, 2021. URL https://www.who.int/publications/i/item/9789240028173.

[303] WHO. Who announces updated definitions of extensively drug-resistant tuberculosis, 2021.

[304] Joanne M Willey, Linda Sherwood, and Christopher J Woolverton. *Prescott's microbiology*. McGraw-Hill, New York, NY, 2014. ISBN 9780073402406 0073402400.

[305] A.B. Williams. Genome Instability in Bacteria. In *Genome Stability*, pages 69–85. Elsevier, 2016. doi : 10.1016/B978-0-12-803309-8.00005-7. URL https://linkinghub.elsevier.com/retrieve/pii/B9780128033098000057.

[306] C. R. Woese and G. E. Fox. Phylogenetic structure of the prokaryotic domain : The primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America*, 1977. ISSN 00278424. doi : 10.1073/pnas.74.11.5088.

[307] WOH. Global tuberculosis report. Technical report, 2023.

[308] Tingting Yang, Mingyu Gan, Qingyun Liu, Wenying Liang, Qiqin Tang, Geyang Luo, Tianyu Zuo, Yongchao Guo, Chuangyue Hong, Qibing Li, et al. Sam-tb : a whole genome sequencing data analysis website for detection of mycobacterium tuberculosis drug resistance and transmission. *Briefings in bioinformatics*, 23(2) : bbac030, 2022.

[309] Yinfei Yang, Ning Jin, Kuo Lin, Mandy Guo, and Daniel Cer. Neural retrieval for question answering with cross-attention supervised data augmentation. 9 2020.

[310] Qing-Qin Yin, Hai-Can Liu, Wei-Wei Jiao, Qin-Jing Li, Rui Han, Jian-Ling Tian, Zhi-Guang Liu, Xiu-Qin Zhao, Ying-Jia Li, Kang-Lin Wan, A-Dong Shen, and Igor Mokrousov. Evolutionary history and ongoing transmission of phylogenetic sublineages of mycobacterium tuberculosis beijing genotype in china. *Sci. Rep.*, 6(1) : 34353, September 2016.

[311] Eiji Yokoyama, Yushi Hachisu, Ruiko Hashimoto, and Kazunori Kishida. Concordance of variable-number tandem repeat (VNTR) and large sequence polymorphism (LSP) analyses of mycobacterium tuberculosis strains. *Infect. Genet. Evol.*, 10(7) : 913–918, October 2010.

[312] Eiji Yokoyama, Yushi Hachisu, Ruiko Hashimoto, and Kazunori Kishida. Population genetic analysis of mycobacterium tuberculosis beijing subgroup strains. *Infect. Genet. Evol.*, 12(4) :630–636, June 2012.

[313] Eiji Yokoyama, Yushi Hachisu, Tomotada Iwamoto, Noriko Nakanishi, Kentaro Arikawa, Takayuki Wada, Junji Seto, and Kazunori Kishida. Comparative analysis of mycobacterium tuberculosis beijing strains isolated in three remote areas of japan. *Infect. Genet. Evol.*, 34 :444–449, August 2015.

[314] Rima Zein-Eddine, Guislaine Refrégier, Jorge Cervantes, and Noemí Kaoru Yokobori. The future of crispr in mycobacterium tuberculosis infection. *Journal of biomedical science*, 30 :34, 5 2023. ISSN 1423-0127. doi : 10.1186/s12929-023-00932-4.

[315] Zhemin Zhou, Nabil-Fareed Alikhan, Khaled Mohamed, Yulei Fan, Agama Study Group, and Mark Achtman. The enterobase user's guide, with case studies on salmonella transmissions, yersinia pestis phylogeny, and escherichia core genomic diversity. *Genome research*, 30 :138–152, 1 2020. ISSN 1549-5469. doi : 10.1101/gr.251678.119.

[316] M R Zimmerman. Pulmonary and osseous tuberculosis in an egyptian mummy. *Bulletin of the New York Academy of Medicine*, 55 :604–8, 6 1979. ISSN 0028-7091.

[317] Cristina Kraemer Zimpel, José Salvatore L. Patané, Aureliano Coelho Proença Guedes, Robson F. de Souza, Taiana T. Silva-Pereira, Naila C. Soler Camargo, Antônio F. de Souza Filho, Cássia Y. Ikuta, José Soares Ferreira Neto, João Carlos Setubal, Marcos Bryan Heinemann, and Ana Marcia Sa Guimaraes. Global Distribution and Evolution of Mycobacterium bovis Lineages. *Frontiers in Microbiology*, 11, may 2020. ISSN 1664-302X. doi : 10.3389/fmicb.2020.00843. URL https://www.frontiersin.org/article/10.3389/fmicb.2020.00843/full.

Titre : Exploiter l'intégralité des données de séquençage de Mycobacterium tuberculosis: une plateforme pour l'analyse génomique in-silico à grande échelle

Mots clés : Mycobacterium tuberculosis, in-silico, séquençage

Résumé : La tuberculose demeure un problème de santé publique majeur à l'échelle mondiale. Selon l'Organisation mondiale de la santé, en 2022, la tuberculose était la deuxième cause de décès dans le monde due à un seul agent infectieux. Depuis plus d'une décennie, l'étude de cette maladie peut également se faire par l'analyse des génomes obtenus de nombreuses souches de tuberculose. Avec le développement des méthodes de séquençage et la réduction de leur coût, on compte actuellement plus de 160 000 séquences de génomes disponibles publiquement. Il s'agit d'une source de données d'une grande richesse mais qui reste une source de données brute, qu'il reste à analyser. Cette thèse introduit la plateforme TB-Annotator qui a permis l'analyse de l'intégralité des données de séquençage de *Mycobacterium tuberculosis* publiquement disponibles. Cette plateforme met en oeuvre de nombreuses méthodes in-silico spécialement développées et adaptées à l'extraction de toutes les caractéristiques génomiques propres au MTBC, notamment dans la détection des variations structurales et dans l'optimisation des pipelines bioinformatiques. Les capacités d'analyse en temps réel de cette masse d'information par TB-Annotator ont permis plusieurs découvertes pour la phylogénie et la phylogéographie du MTBC. Cette thèse décrit notamment la découverte d'une nouvelle lignée rare, la lignée 10, mise en évidence par 2 séquences publiques non classifiées après analyse des 160 000 séquences disponibles. Plusieurs améliorations de la définition de lignées et sous-lignées sont aussi décrites, notamment L2, L4, L6 et L7. L'ouverture de cette plateforme devrait permettre de simplifier l'accessibilité des analyses à grande échelle pour l'épidémiologie, la phylogénétique, la phylogéographie et la résistance aux médicaments.

Title : Leveraging the entire sequencing data of Mycobacterium tuberculosis: a platform for large-scale in-silico genomic analysis.

Keywords : Mycobacterium tuberculosis, in-silico, sequencing

Abstract : Tuberculosis remains a major public health issue worldwide. According to the World Health Organization, in 2022, tuberculosis was the second leading cause of death globally due to a single infectious agent. For over a decade, the study of this disease has also been possible through the analysis of genomes obtained from numerous tuberculosis strains. With the development of sequencing methods and the reduction in their cost, there are currently more than 160,000 publicly available genome sequences. This is a rich data source, but it remains raw data that needs to be analyzed. This thesis introduces the TB-Annotator platform, which enabled the analysis of all publicly available Mycobacterium tuberculosis sequencing data. This platform implements numerous specially developed in-silico methods adapted to extract all genomic features specific to the MTBC, particularly in detecting structural variations and optimizing bioinformatics pipelines. The real-time analysis capabilities of this vast amount of information by TB-Annotator have led to several discoveries regarding the phylogeny and phylogeography of the MTBC. This thesis notably describes the discovery of a new rare lineage, lineage 10, identified through 2 unclassified public sequences after analyzing the 160,000 available sequences. Several improvements in the definition of lineages and sub-lineages are also described, notably L2, L4, L6, and L7. The opening of this platform should simplify the accessibility of large-scale analyses for epidemiology, phylogenetics, phylogeography, and drug resistance.

Université Bourgogne Franche-Comté
32, avenue de l'Observatoire
25000 Besancon