WHOLE-GENOME SEQUENCING FOR EPIDEMIOLOGIC STUDIES OF TUBERCULOSIS

Robyn S. Lee

Department of Epidemiology, Biostatistics and Occupational Health

McGill University, Montréal

April 2016

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy (Ph. D)

ABSTRACT

Between 2011-2012, a single Inuit village of Nunavik had 50 culture-confirmed cases of tuberculosis (TB), representing an incidence of 5% for that year. Among those with recent infection, 20% progressed to active TB disease, compared to the expected attack rate of ~5%. Classical molecular typing methods suggested this was a single, point-source outbreak, alarming in magnitude to both public health and the community. However, previous molecular studies in the Arctic had revealed limited bacterial diversity. Therefore, it was also possible these methods simply lacked sufficient resolution to discriminate true transmission events from the absence of transmission. Recent work using a newer method, whole genome sequencing (WGS), suggests this technique provides higher resolution genotyping data, and thus may more accurately discriminate between transmission and reactivation of remote infection. However, while WGS is poised to become the leading method for detecting infectious disease transmission, bioinformatics pipelines – steps required to covert raw genetic information into useable data for epidemiology – lack standardization, with the potential risk that arbitrary decisions during data processing may affect epidemiologic inferences. Through this thesis, I have applied WGS to an outbreak, and then a population-based study, to evaluate TB transmission in a high-incidence setting. In so doing, I have examined both methodological aspects and clinical applications of WGS to TB control.

RESUMÉ

Entre 2011 et 2012, 50 cas de tuberculose confirmés par culture ont été recensés dans un village Inuit de Nunavik, ce qui représente une incidence de 5% pour cette période. Parmi ceux ayant été infectés récemment, 20% ont développé une tuberculose active, alors que l'on s'attendait à ~5%. Des méthodes classiques de typage ont suggéré qu'il s'agissait d'une éclosion provenant d'une unique source ponctuelle et alarmante à la fois d'un point de vue de santé publique et pour la communauté. Cependant, des études moléculaires dans la région Arctique ont révélé une diversité bactérienne limitée. Il était donc possible que ces méthodes manquaient tout simplement de résolution afin de discriminer de réels événements de transmission d'une absence de transmission. De récents travaux utilisant une nouvelle méthode, le séquençage de génomes entiers, suggèrent que cette technique fournit des données de meilleure résolution et permettrait de discriminer de manière précise la transmission de la réactivation d'une infection précédente. Cependant, tandis que le séquençage de génomes devient la méthode de choix pour la détection de transmission de maladies infectieuses, les pipelines bio-informatiques – étapes requises permettant de traiter les données génétiques brutes en informations utilisables en épidémiologie – manquent de standardisation, avec le risque potentiel que le traitement arbitraire des données puisse affecter les conclusions épidémiologiques. À travers cette thèse, j'ai appliqué le séquençage de génomes entiers à une éclosion, puis une étude des populations, afin d'évaluer la transmission de la tuberculose dans le cadre d'une grande incidence. En procédant ainsi, j'ai examiné à la fois les aspects méthodologiques et les applications cliniques du séquençage de génomes entiers dédiés au contrôle de la tuberculose.

ACKNOWLEDGEMENTS

There are a number of individuals without whom I could not have completed this work. While the detailed and specific contributions authors have made to each manuscript are listed in the following section, "Contributions of Authors", I would like to particularly acknowledge the following persons:

My supervisor, Dr. Marcel Behr, has been an amazing support throughout this endeavour. He has constantly encouraged me and believed in my capabilities, while simultaneously pushing me to be better. I have benefited so much from his expertise, his critical feedback of my work and the learning opportunities he has provided throughout my PhD. He taught me what it means to be a scientist, including how to think critically and develop a question from start to finish. Any future success I experience as a researcher will be in part because of him and the skills he has taught me. I could not have been luckier with respect to a mentor.

Dr. Jean-Francois Proulx, a collaborator on many of the manuscripts herein, has provided invaluable feedback on this work. I have benefited greatly from his experience in the North and appreciated his willingness to provide input on drafts at various stages of development. I also want to acknowledge Dr. Proulx's years of work collecting and maintaining epidemiological records in Nunavik; without his attention to detail, these projects would not have been possible.

This work could also not have been done without the input and support of the council of the village in Nunavik. I would also like to acknowledge the employees of the Centre Local de Services Communautaires from the village, as well as the staff at the Centre de Santé de Tulattavik de l'Ungava and Nunavik Regional Board of Health and Social Services, who worked tirelessly during the outbreak.

I would also like to thank my thesis committee members, Drs. Harper and Benedetti, for their critical review of my work over the past years and for always being willing to trek out to "remote" sites for my committee meetings.

I would also like to thank Dr. Menzies who was on my thesis committee, but was also the person who first involved me in this project. Due to his clinical involvement in Nunavik, Dr. Menzies introduced me to staff at the Centre de Santé Tulattavik de l'Ungava. Within 10 days, I was up in the Arctic working as a nurse, helping with the public health response to the outbreak. This experience changed my life.

Finally, I want to thank my brother for his consistent support and my good friends (and former/current epidemiology students) Christine Sabapathy and Ania Syrowatka, for listening patiently and providing much needed humour throughout this process.

Statement of financial support

I would like to acknowledge and thank the Canadian Institutes of Health Research for the funding of this research, in the form of an Operating Grant awarded to Drs. Behr and Menzies. This grant also funded my stipend for several years of my PhD. In addition, I would like to thank my department for awarding me a Graduate Excellence Fellowship, as well as the Research Institute of the McGill University Health Centre, which awarded me two years of fellowship support.

PREFACE. CONTRIBUTIONS OF AUTHORS

Manuscript I (Published)

Lee RS, Radomski N, Proulx J-F, Manry J, McIntosh F, Desjardins F, Soualhine H, Domenech P, Reed MB, Menzies D, Behr MA. Reemergence and re-amplification of tuberculosis in the Canadian Arctic. *J Infect Dis* 2015;211(12):1905-1914

Robyn S. Lee (Ph. D Candidate):

Collected epidemiological and clinical data through chart review and patient interview in clinical capacity for 2011-2012, and compiled these into a database for Nunavik Regional Board of Health and Social Services. Performed phylogenetics and combined epidemiological data with genetic to identify transmission networks during the 'outbreak'. Performed IS*6110* restriction fragment length polymorphism of study isolates included in manuscript. Conducted all statistical and epidemiological analyses. Contributed to study design (helped write CIHR operating grant), helped determine objectives of analysis and interpret results. Wrote the first draft of the manuscript.

Nicolas Radomski (former Post-doctoral Fellow), The Research Institute of McGill University Health Centre:

Developed bioinformatics pipeline used in this manuscript, with alignments provided by the McGill University and Génome Québec Innovation Centre. Performed Sanger confirmation of selected single-nucleotide polymorphisms (SNPs) and preliminary phylogenetics. Contributed to study design (helped write CIHR operating grant), interpretation of the data and helped write the manuscript.

Jean-Francois Proulx (MD), Nunavik Regional Board of Health and Social Services:

Collected and validated epidemiological data for the villages of Nunavik for the years of study. Helped develop study objectives, interpret results and provided critical feedback on the manuscript.

Jeremy Manry (Post-doctoral Fellow), Departments of Human Genetics and Medicine, McGill University:
Helped develop the bioinformatics pipeline and provided critical feedback on the manuscript.

Fiona McIntosh (Laboratory technician for Dr. Behr), The Research Institute of McGill University Health Centre:
Performed DNA extractions and provided critical feedback on the manuscript.

Francine Desjardins (Laboratory technician, retired), Mycobacteriology laboratory of the McGill University Health Centre:
Received and grew all *Mycobacterium tuberculosis* samples and provided critical feedback on the manuscript.

Hafid Soualhine (Ph. D), Laboratoire de Santé Publique du Québec:
Provided mycobacterial interspersed repetitive unit data for *Mycobacterium tuberculosis* isolates from patients, as well as drug susceptibility testing results. Performed DNA extractions.

Pilar Domenech (Research Associate, Laboratory of Dr. Reed), The Research Institute of McGill University Health Centre:
Helped develop the bioinformatics pipeline and provided critical feedback on the manuscript.

Michael Reed (Ph. D), Department of Microbiology, McGill University and The Research Institute of McGill University Health Centre:
Helped develop the bioinformatics pipeline and provided critical feedback on the manuscript.

Dick Menzies (MD MSc), Respiratory Epidemiology and Clinical Research Unit, Montreal Chest Institute and The Research Institute of McGill University Health Centre:
Co-principal investigator. Helped design the study and obtained funding via CIHR. Contributed to writing of the manuscript and provided feedback on results.

Marcel A. Behr (MD MSc), Microbiologist-in-Chief, McGill University Health Centre, Department of Medicine, McGill University and The Research Institute of McGill University Health Centre:

Co-principal investigator. Designed and supervised the study. Obtained funding via CIHR and identified key collaborators. Determined objectives of analysis. Interpreted results and wrote first draft of the manuscript and provided critical feedback on the work.

Manuscript II (Published)

Lee RS, Radomski N, Proulx J-F, Levade I, Shapiro BJ, McIntosh F, Soualhine H, Menzies D, Behr MA. Population genomics of *Mycobacterium tuberculosis* in the Inuit. *Proc Natl Acad Sci USA* 2015;112(44):13609-13614

Robyn S. Lee, Ph. D Candidate:

Conducted review of historical literature. Performed final SNP annotation, analyses excluding repetitive regions, model selection and phylogenetic trees used in manuscript. Conducted all statistical and epidemiological analyses, including evaluation of transmission, dN/dS calculations and descriptive statistics. Performed Bayesian molecular dating analyses. Helped determine objectives of analysis, interpreted results and wrote the manuscript.

Nicolas Radomski (former Post-doctoral Fellow), The Research Institute of McGill University Health Centre:

Conducted literature review of gene categories for tuberculosis. Developed bioinformatics pipeline, with alignments to the H37Rv reference genome provided by the McGill University and Génome Québec Innovation Centre. Performed alignments of reads to the CDC1551 reference genome and SNP calling. Identified and performed polymerase chain reaction (PCR) confirmation of large deletions. Performed preliminary phylogenetics. Helped interpret results and helped write the manuscript.

Jean-Francois Proulx (MD), Nunavik Regional Board of Health and Social Services:

Collected and validated epidemiological data for the villages of Nunavik for the years of study. Helped develop study objectives, interpret results and provided critical feedback on the manuscript.

Ines Levade (Ph. D Candidate), Department of Biology, Université de Montréal:

Performed Bayesian molecular dating analyses and generated skyline plots. Helped interpret the data and critically reviewed the manuscript.

B. Jesse Shapiro (Ph. D), Department of Biology, Université de Montréal:
Helped inform evolutionary analyses and interpret results. Contributed to writing of the manuscript and provided valuable feedback on content.

Fiona McIntosh (Laboratory technician for Dr. Behr), The Research Institute of McGill University Health Centre:
Performed DNA extractions and provided critical feedback on the manuscript.

Hafid Soualhine (Ph. D), Laboratoire de Santé Publique du Québec:
Provided mycobacterial interspersed repetitive unit data for *Mycobacterium tuberculosis* isolates from patients, as well as drug susceptibility testing results. Performed DNA extractions.

Dick Menzies (MD MSc), Respiratory Epidemiology and Clinical Research Unit, Montreal Chest Institute and The Research Institute of McGill University Health Centre:
Co-principal investigator. Helped design the study and obtained funding via CIHR. Contributed to writing of the manuscript and provided feedback on results.

Marcel A. Behr (MD MSc), Microbiologist-in-Chief, McGill University Health Centre, Department of Medicine, McGill University and The Research Institute of McGill University Health Centre:
Co-principal investigator. Designed and supervised the study. Obtained funding via CIHR and identified key collaborators. Determined objectives of analysis, imputed ancestral sequences, interpreted the data and wrote the manuscript.

Manuscript III (In review)

Lee RS, Proulx J-F, Menzies D, Behr MA. Progression to tuberculosis disease increases with multiple exposures. *Eur Respir J.*

The results of this manuscript were presented internationally at a Keystone Symposia, Keystone, Colorado in February 2016.

Robyn S. Lee, Ph. D Candidate:
Collected epidemiological and clinical data and compiled these into a database as previous. Helped in design of the study, determined objectives of analysis, performed statistical analyses and wrote first draft of the manuscript.

Jean-Francois Proulx (MD), Nunavik Regional Board of Health and Social Services:
Collected and validated epidemiological data for the villages of Nunavik for the years of study. Helped develop study objectives, interpret results and provided critical feedback on the manuscript.

Dick Menzies (MD MSc), Respiratory Epidemiology and Clinical Research Unit, Montreal Chest Institute and The Research Institute of McGill University Health Centre:
Co-principal investigator. Designed the study, helped interpret results, and provided critical feedback on the manuscript.

Marcel A. Behr (MD MSc), Microbiologist-in-Chief, McGill University Health Centre, Department of Medicine, McGill University and The Research Institute of McGill University Health Centre:
Co-principal investigator. Designed and supervised the study. Helped determine objectives of analysis, interpreted results and helped write the manuscript.

Manuscript IV (Accepted April 5, 2016)

Lee RS and Behr MA. Does choice matter? Reference-based alignment for molecular epidemiology of tuberculosis. *J Clin Micro*

Robyn S. Lee, Ph. D Candidate:

Conceived and designed the research, developed the bioinformatics pipeline and performed all bioinformatics, statistical and epidemiological analyses. Interpreted results and wrote the manuscript.

Marcel A. Behr (MD MSc), Microbiologist-in-Chief, McGill University Health Centre, Department of Medicine, McGill University and The Research Institute of McGill University Health Centre:

Designed the research, supervised the study, interpreted results and contributed to writing the manuscript.

Manuscript V (Published)

Lee RS and Behr MA. The implications of whole genome sequencing in the control of tuberculosis. 2016. *Ther Adv Infect Dis*;3(2):47-62.

Robyn S. Lee, Ph. D Candidate:

Conducted the literature review and abstracted the data. Produced the figures and wrote the manuscript.

Marcel A. Behr (MD MSc), Microbiologist-in-Chief, McGill University Health Centre, Department of Medicine, McGill University and The Research Institute of McGill University Health Centre:

Review was commissioned to Dr. Behr. Wrote the manuscript and helped interpret the data from a clinical perspective.

STATEMENT OF ORIGINALITY

This thesis represents an original contribution to the literature. It is the first to apply whole genome sequencing (WGS) to the Canadian North and in doing so, has greatly increased our understanding of TB transmission in this context, where classical molecular typing methods were insufficient to discern recent transmission from reactivation of remote infection. This is the first study to demonstrate the importance of local strain diversity in establishing thresholds of single nucleotide polymorphisms for delineating transmission, and has illustrated the importance of using WGS in combination with epidemiologic data for resolution of TB epidemics. In addition, this work has investigated an important bioinformatics decision in the analysis of WGS data; as these may ultimately influence our epidemiologic inferences, such validation is critical. Finally, from a substantive perspective, our finding that the predominant strain of TB in Nunavik has undergone relaxation of purifying selection challenges an emerging view in TB that epidemiologic success depends on characteristics of the strain. Instead, this suggests social and/or environmental factors have facilitated propagation of TB in the Inuit population.

This thesis is presented in manuscript form, and is comprised of 5 articles: 3 published, 1 accepted, and 1 currently undergoing peer review. In addition to these publications, parts of this work have been presented in various formats at the following external conferences: The Union Against Tuberculosis and Lung Disease, North America Region (Boston, Massachusetts 2014, Vancouver, British Columbia 2015 and Denver, Colorado 2016) and the Keystone Symposia on Tuberculosis Co-morbidities and Immunopathogenesis (Keystone, Colorado 2016).

TABLE OF CONTENTS

CHAPTER 4. MATERIALS AND METHODS

FIGURES

TABLES

CHAPTER 1. INTRODUCTION

1.1 Background and rationale

In November of 2011, there were two cases of tuberculosis diagnosed in a small Arctic community in Nunavik, Québec (**Figure 1-1** and **1-2**). Despite local efforts to reduce transmission, by March of 2012, this had escalated to 17 cases. A massive public health response was initiated, resulting in more than 2/3 of the community being investigated for contact. Ultimately, a total of 50 individuals were diagnosed with microbiologically-confirmed (i.e., culture-positive) disease in this community, representing an incidence of over 5% for that year. This event was also characterized by an extraordinarily high attack rate. In contrast to the commonly-cited 2-5% which are expected to progress within the years immediately following infection, approximately 20% of those with recent infection progressed rapidly to disease in this village. It was unclear what had led to such an epidemic of TB in this community or the high rate of progression, as there was low HIV prevalence, negligible drug resistance and high reported adherence to latent TB prophylaxis. Contact investigations revealed a tangled web of epidemiological links between cases, with most having multiple potential sources of transmission. The complexity of these epidemiologic links made it difficult to resolve transmission using these data alone. Thus, molecular typing methods were considered.

Since Insertion Sequence *6110* (IS*6110*) restriction fragment length polymorphism (RFLP, see **Appendix 1** for a glossary of molecular epidemiology-related terms and acronyms) was first proposed as an epidemiologic tool in 1991 (1), molecular genotyping methods have assumed a critical role in delineating tuberculosis transmission. Using these tools, pairs of bacterial samples ('isolates') from patients are considered the result of recent transmission if they share an identical or highly similar fingerprint. Conversely, isolates with different fingerprints are thought to represent independent progression of disease, without a transmission link between them. When isolates have a unique fingerprint, not matching another in the database, this isolate is inferred to be the result of reactivation of an infection acquired in a different time (in the past) or place (in a different country). Examining such data in the context of this Arctic 'outbreak' revealed identical patterns for 49/50 isolates, suggestive of a single outbreak. However, previous work in this region had shown limited bacterial diversity (2), therefore

shared ancestry could provide an alternative explanation for the lack of variation in RFLP. A third possible explanation, which could not be evaluated using a low resolution typing modality, such as RFLP, would be both limited genetic variability and some recent transmission.

At the onset of this work, a single study had applied a newer method – whole genome sequencing (WGS) – to investigate a TB outbreak in British Columbia (3). Unlike classical genotyping methods, which examine only ~1% of the *Mycobacterium tuberculosis* genome, WGS interrogates all 4.4 million base-pairs. WGS predominantly aims to identify and quantify single base changes in the genome compared to a reference, called 'single nucleotide polymorphisms' (SNPs). These SNPs can then be used to compare isolates for epidemiology. In this seminal paper, authors illustrated the potential of WGS to resolve outbreaks beyond the capabilities of classical methods. Based on this work, we ultimately applied this new method to investigate transmission in the North, to further resolve this unique event.

This thesis is largely comprised of our epidemiological investigations of TB in the Arctic. However, it is important to note that large-scale WGS, such as that needed for outbreak investigation, has only recently become feasible due to restrictions both in technology and cost. As such, the analytic approaches largely lack standardization. When raw data is produced by sequencing platforms, for example, there are a number of processing steps (referred to as a 'bioinformatics pipeline') required to produce the final dataset we use in genomic epidemiology, with many different software choices available for each. As I learned to work with WGS data, questions arose about the analytic choices involved as well as future applications of this tool to TB.

The results of this work are presented in the form of a manuscript-based thesis, comprised of 5 articles (3 published, 1 accepted, and 1 currently under review). Please note that the references for each of these manuscripts are retained and distinct from the overall reference list for Chapters 1-4 and 10, which represent the Introduction, Reviews of the literature (2), Materials and Methods, and the Discussion.

The detailed rational for each study included in this thesis is as follows:

Manuscript I, "Reemergence and re-amplification of tuberculosis in the Canadian Arctic" is presented in **Chapter 5**. In this study, WGS was applied in conjunction with field and clinical epidemiology to understand transmission during the outbreak. Through this in-depth analysis, it was discovered that this singular 'outbreak' was actually comprised of at least 6 different groups of transmission. As only 20% of non-household contacts with active TB shared the same genotype as their putative source, this also indicated that contact investigation beyond the household was of limited utility in this context. As a direct consequence of these findings, the regional public health unit has initiated village-wide screenings in lieu of extended contact investigation during more recent surges of TB in the North. In addition to these substantive findings, this was the first study to demonstrate the importance of local strain diversity in ruling out transmission using WGS; our validation study revealed that even patients without epidemiologic links *and* from different villages could have isolates with as few 2 SNPs between them. Finally, to our knowledge, this was also the first study to demonstrate clonal replacement (wherein one strain was completely replaced by another over time), suggesting that it *is* feasible to eradicate strains of TB in this context.

Manuscript II, "Population genomics of *Mycobacterium tuberculosis* in the Inuit" is presented in **Chapter 6** and expands on the first study to include all villages of Nunavik. This study was conducted in response to village and public health concerns that a new, hyper-virulent strain of tuberculosis had arrived in the Canadian North. This is the first population-based study to use WGS to examine transmission in the Canadian Arctic, refuting our own lab's previous interpretation that suggested that there was ongoing transmission between villages of the Nunavik (2). The finding that the predominant circulating strain of TB was introduced into the region a century ago, with no evidence of increased virulence, suggests that clinical management of TB in this region can continue as per usual. As TB has undergone genome-wide relaxation of purifying selection since this time, this also challenges an emerging view in TB research that the epidemiological success of TB likely depends on underlying virulence of the bacteria. Instead, our data suggests that social and environmental conditions favourable to transmission may drive TB in this context, independent of bacterial factors.

Manuscript III, "Progression to tuberculosis disease increases with multiple exposures" is found in **Chapter 7** and addresses another public health concern, pertaining to the extraordinary attack rate in the 'outbreak' village. This work was inspired by the observation that numerous cases had contact with multiple potential sources, and simultaneously, with multiple different genotypes as identified in Manuscript I. Previous studies have suggested a potential role for infectious inoculum in progression to disease, using proxy measures such as close versus casual contact (4) or occupational exposure (5). This dichotomous approach may result in substantial residual confounding; thus far, no studies have utilized a quantitative exposure metric. While many cite a 2-5% risk of progression within the first 5 years after infection (6-9), this study suggests that such a single estimate of risk may under-estimate the risk of progression in high-incidence settings, where multiple exposures frequently occur. Thus, epidemiologic models and clinical assessment of risk of progression should potentially take the intensity of exposure into consideration.

Overall, these three studies, along with two additional case-controls not included in this thesis (10, 11), represent the first comprehensive analysis to apply genomic, classical (via case-control design) and boots-on-the-ground field epidemiology to a single TB 'outbreak'.

Manuscript IV in **Chapter 8**, "Does choice matter? Reference-based alignment for molecular epidemiology of tuberculosis" was developed based on methodological questions that arose during the analysis of WGS data. In developing my bioinformatics pipeline for *M. tuberculosis*, one consideration was the choice of reference genome. It had been suggested that the use of a reference genome from a different lineage than the isolates under study could result in substantial loss of data, as sequenced 'reads' cannot be aligned to loci that are absent from the reference. However, high-quality, assembled reference genomes are not readily available for all lineages of TB. Therefore, we examined the impact of such a decision on phylogenetic trees and estimates of transmission in an epidemiologic 'outbreak'. Results of this study can easily be extrapolated from this unique low-diversity dataset to other environments with high strain variability, setting the standard for reference-based analyses of tuberculosis.

During the past few years, it has become largely accepted that WGS is the new gold standard for genomic epidemiology, despite the remaining questions in bioinformatics analysis. My final manuscript (Manuscript V, in **Chapter 9**), "The implications of whole genome sequencing for control of tuberculosis" goes beyond WGS as a tool for epidemiology to consider a new application: its potential utility for clinical diagnostics and prediction of drug resistance. While this new use of WGS data has recently garnered much interest from both researchers and public health departments alike, reviewing the available literature suggests that such applications of WGS for tuberculosis, while promising, would presently be premature. It cautions the need for ongoing phenotypic drug susceptibility testing until more information is available on resistance-conferring mutations, an important message as some countries are considering eliminating other diagnostics and phenotypic drug susceptibility testing from their clinical workflow (e.g., (12)).

This work was done in collaboration with the Nunavik Regional Board of Health and Social Services, and the village council of the community in Nunavik. Manuscripts I through III aim to address important questions and concerns expressed by these parties. Results have been disseminated back to the community and have ultimately helped inform public health interventions in this region. In utilizing WGS for this research, we have not only contributed to our understanding of TB transmission in the Canadian North but also addressed important and timely issues in the use of such data. As WGS has also become the gold standard for epidemiology of other pathogens, this work not only has implications for TB, but other infectious diseases as well.

FIGURE 1-1. Map of Canada. Nunavik is the Arctic region of Québec, indicated in red. Source of image: fr.wikepedia.org.

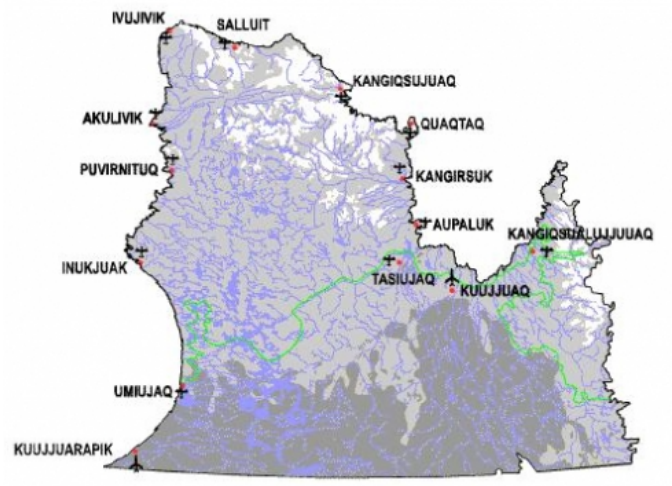


FIGURE 1-2. Villages of Nunavik. All 14 communities of Nunavik are indicated. There are no roads connecting these communities. Most of the year, travel between these villages is restricted to small plane, with flights to and from Montréal connecting via Kuujjuaq. This region has two hospitals: one in Purvinituq, which services villages on the Hudson coast (west), and one in Kuujjuaq, which services villages on the Ungava coast (east). The remaining villages have full-time nursing stations. Source of image: https://www.mcgill.ca/hssaccess/trhpp/m1program/projects/17-nunavik

1.2 Objectives

The overall aim of this thesis is to explore the use of WGS for the epidemiology of TB, increase understanding of transmission in the Canadian Arctic, and investigate methodological aspects of WGS data analysis. Specific objectives include:

1. To resolve transmission in a major TB outbreak in Northern Quebec using WGS and clinical epidemiologic data (Manuscript I)

2. To examine TB transmission within and across the villages of Nunavik, in order to address public health concerns about a new, hyper-virulent strain entering this region (Manuscript II)

3. To examine risk factors, including exposure to different genotypes of *M. tuberculosis* as identified by WGS, for progression to active TB during the same outbreak (Manuscript III)

4. To examine the effect of using different reference genomes on epidemiologic inferences (Manuscript IV)

5. To evaluate the feasibility of WGS moving from an epidemiologic tool to a diagnostic test, and its potential to affect clinical management decisions (Manuscript V)

CHAPTER 2. REVIEW OF THE LITERATURE PART 1 - TUBERCULOSIS

2.1 Pathogenesis

2.1.1 Etiology

Tuberculosis is a disease caused by the bacterium *Mycobacterium tuberculosis*, in which humans are the only known reservoir. Pulmonary tuberculosis is the most frequent and most transmissible form of the disease (6). When a person with active pulmonary TB coughs, speaks or sings (13, 14), droplet nuclei 1-5 μm in size containing *M. tuberculosis* become suspended in the air, where they can remain for minutes to hours (15). When another individual inhales these droplets, *M. tuberculosis* passes into the lower respiratory tract and terminal alveoli, where infection may become established (16). Among immunocompetent hosts, it is often cited that 2-5% will progress to active TB disease within first few years immediately following this infection ('primary progressive disease'). The remainder enter an asymptomatic state called 'latent tuberculosis infection' (LTBI), with an additional lifetime risk of progression of 5% (6).

2.1.2 Risk factors for tuberculosis

Exposure to *M. tuberculosis* is a necessary but not a sufficient cause of infection. Whether an individual actually develops infection and progresses to disease may depend on several factors, including the characteristics of the host, the source case, the environment and the bacteria.

2.1.2.1 The host

From the host perspective, the innate immune response is thought to play a critical role in preventing infection. It has been shown that establishment of infection is dependent on the ability of *M. tuberculosis* to enter host alveolar macrophages, a form of white blood cell, and evade the host immune system by preventing the fusion of phagosomes with lysosomes, altering anti-microbial effectors and subverting mechanisms that would usually result in programmed cell death (16). Humans with mutations in genes affecting the interferon-ϒ pathway, which is involved in priming macrophages (and T cells) in early infection, are not only more susceptible to tuberculosis but also highly susceptible to infection even by weakly virulent mycobacterium (17). Similarly, deficiencies in pattern-recognition receptors that would

usually detect mycobacteria and trigger signalling cascade leading to macrophage activation, have been linked to infection with *M. tuberculosis* and progression to disease (18).

In addition to genetic factors, host immunity to *M. tuberculosis* infection may be mediated by other epidemiological characteristics. A recent meta-analysis showed that cigarette smoking was associated with 1.8-fold higher odds of LTBI (95% CI 1.5-2.2) and 1.49-2.87-fold higher odds of active TB disease compared to non-smokers (19). The physiological effects of smoking on the immune system are reviewed here (20). In brief, smoking impairs the mucociliary escalator (wherein microbes are trapped in mucous and subsequently moved up and out of the respiratory tract), potentially allowing increased *M. tuberculosis* bacilli to enter the lung. Smoking also drastically increases the number of alveolar macrophages, the target cells of *M. tuberculosis* infection. In addition, smoking impairs the production of pro-inflammatory cytokines by these macrophages when they are infected with *M. tuberculosis,* leading to decreased bacterial clearance (21, 22). Alcohol consumption can also mediate host immunity (23), and has been associated with both increased risk of infection and progression to disease (24). While many suggest this link is causal (24, 25), it is possible that these associations are confounded by increased exposure to *M. tuberculosis* (26).

HIV infection is also an important factor. There is little epidemiologic evidence suggesting HIV increases the risk of infection (27), however, persons who are co-infected with TB and HIV are at much greater risk of progression to TB disease (28). The annual risk of progression to TB disease among those with HIV is estimated at ~10% (29) compared to a lifetime risk of 10% in the HIV-negative population. Not surprisingly, other diseases associated with immune suppression, such as diabetes or chronic kidney disease have also been associated with TB (30, 31).

While it is thought nutritional status can influence host immunity, a role for this in infection in unclear (16). While several studies have shown lower vitamin D_3 levels in subjects with latent TB infection versus those without (32-34), for example, these were not consistent and flawed in design or analysis. Similarly, the impact (if any) of micronutrient status on progression to active TB is currently uncertain (16).

Finally, age is associated with active TB. As age and opportunity for exposure increases, the prevalence of latent tuberculosis infection increases. Children, particularly those infected under 2 years of age (35) are at higher risk of progression to disease. The elderly are also at elevated risk compared to other age groups, due to declining immune function with age ('immunosenescence') (36).

2.1.2.2 The source

A source case is an individual who successfully transmits *M. tuberculosis*, resulting in infection. Pulmonary and laryngeal TB are the most contagious forms of this disease; other forms of extra-pulmonary TB are usually only transmitted by aerosolization of bacteria found in abscesses (13). The presence of bacteria in expectorated sputum has been associated with contagiousness of the source case. A person is classified as having 'sputum smear positive' disease if bacilli are visible under a microscope after staining with either an acid-fast fluorochrome dye (such as auramine O, for fluorescence microscopy) or carbolfuschin acid-fast stains (such as Ziehl-Neelsen, for conventional microscopy) (37). Conversely, a person is said to have 'smear negative' disease when such bacilli are not visible. Persons with sputum smear positive pulmonary TB can expectorate as much as 10^6 to 10^7 acid-fast bacilli per mL of sputum per day (38) and it has been estimated that an untreated sputum smear positive TB case infects 10 persons per year (39). While persons with smear negative disease expel lower amounts of bacteria, at $<10^3$ bacilli per mL, they are also contagious - albeit to a lesser degree (40).

Transmission has also been associated with a cavity on chest x-ray (41). As pulmonary tuberculosis becomes more advanced, necrosis of lung tissue can result in formation of large, air-filled spaces known as cavities. These cavities harbor large numbers of bacilli compared to other lung tissue; a single cavity can contain 10^8 bacteria, while a person with extensive disease can have as much as 10^{12} (6).

The presence of cough has also been suggested as a key factor in transmission (6), though there is surprisingly not much supporting evidence (reviewed in detail in (42)). In 1967, a small study evaluated nocturnal cough frequency in 63 patients. Increased cough was associated with

increased tuberculin skin test (TST) positivity in household contacts under the age of 15, however this was not statistically significant (p=0.11). In this study, sputum smear had a stronger association with infection (p<0.01) (43). Another more recent study has investigated the force of (voluntary) cough in tuberculosis patients; stronger cough and increasing sputum smear grade (i.e., greater numbers of bacilli on microscopy) were associated with having aerosol-based cultures positive for *M. tuberculosis,* but neither was included in final multivariate analyses (44).

Children are generally considered less contagious as they are thought to have predominantly paucibacilliary disease (caused by very few bacteria). However, transmission can rarely occur (e.g., (45)).

2.1.2.3 The environment

Adequate ventilation reduces the amount of droplet nuclei in the air, thereby decreasing the number of bacteria to which one is exposed. Modelling studies based on real-life data suggest transmission would decrease in prisons and other settings if ventilation were to increase (46, 47), while lower ventilation, as measured in a hospital setting, was associated with increased risk of *M. tuberculosis* infection in health care workers (48).

In addition to ventilation, the proximity to the source case plays a role in infection. In an outbreak on a naval ship in the 1960s, crew members who resided in the same quarters as individuals with active TB disease were more likely to develop infection and progress to active TB disease once infected compared to those in other compartments, even with shared ventilation (49). A recent meta-analysis continues to support these results, showing that infection occurred in ~30-50% of 'household' and 'close' contacts (though the definitions of these varied widely by study) (27), while the risk was ~10% lower for 'casual' contacts in high-income settings.

2.1.2.4 The bacteria

Seven lineages of *M. tuberculosis* have been identified (50, 51). It has been suggested that different lineages may vary in their ability to cause infection and subsequent disease. This is

partly based on animal studies, which have illustrated increased pathogenicity of Beijing strains (a subset of lineage 2) (52). Such strains have also been shown to have a higher mutation rate and increased frequency of drug resistance mutations *in vitro* (53). While some epidemiological data suggests that this strain or sublineages thereof may contribute to higher rates of transmission compared to others (54-56), these data are not consistent across different populations or geographical regions (57). Many of these studies also rely on detected cases as a measure of transmission, which may reflect increased progression to disease rather than increased propensity to infect. In agreement with the latter, a study in The Gambia showed similar infectivity but decreased progression to active TB disease with *M. africanum* compared to other TB lineages (58).

2.1.3 Preventing progression to active disease

It is currently recommended that persons with LTBI undergo a minimum of 6 months of isoniazid prophylaxis to prevent future progression to active TB disease (59). A 4-month course of rifampin has also been proposed more recently, as an alternative to this treatment (6). In low-incidence countries, the acceptance of LTBI treatment is thought to be a cornerstone of population-level TB control. However, at the level of the patient, the decision to administer LTBI prophylaxis depends on the individual's underlying risk of progression and potential contraindications for treatment (6).

2.2 Epidemiology

2.2.1 The global epidemiology of TB

Despite being a treatable disease, there were 9 million incident cases of TB worldwide in 2013, with 1.1 million deaths in the HIV-negative population (60). Control of tuberculosis is dependent on rapid diagnosis of contagious cases and access to appropriate treatment. However, programmatic limitations make this difficult in many countries of the world. Reliance on sputum smear microscopy for detection of TB in some regions means that numerous smear negative cases who are also capable of infecting others (40) will - at least initially - be missed. Detecting and treating latent TB infection is also beyond the current capabilities of many countries, potentially exacerbating the problem as it is thought that ~10% of these will progress to active TB during the course of their lifetime without prophylaxis. The rise of HIV has also

contributed to the TB epidemic, particularly in African countries. In this region, 34% of all TB cases were HIV co-infected (60). Multi-drug resistant (MDR) TB, defined as resistance to two of the first-line anti-chemotherapeutics, rifampin and isoniazid, is also becoming an increasing concern; MDR-TB treatment is considerably longer than conventional therapy, has more side effects and is often unsuccessful (61). The World Health Organization (WHO) estimated that 480,000 individuals were diagnosed with MDR-TB in 2013 (60). It has recently been proposed that 95.9% of incident MDR cases are in fact due to transmission of an MDR strain (95% uncertainty interval 68.0-99.6) (62), rather than acquired drug resistance (i.e., resistance developed during treatment due to inadequate regimes or poor adherence) as was the previous contention.

2.2.2 Tuberculosis in Canada

Classified as 'low-incidence' by the WHO, Canada is one of 33 countries and territories in the world targeting TB elimination by 2035 (63). Yet, despite an overall rate of 4.4/100,000 in 2014 (64), certain subpopulations continue to experience much higher rates of TB in Canada. Aboriginals and the foreign-born represented 21% and 69% of the TB cases in Canada in 2014 despite only comprising 4% and 22% of the Canadian population, respectively (64). The causes of these high rates of TB are thought to be distinct, with TB in the Aboriginal population primarily due to ongoing transmission and TB in the foreign-born primarily due to reactivation of remote infection from one's country of origin (65). The distinction between ongoing transmission and reactivation is critical, as different public health interventions are required to reduce TB depending on the cause.

2.2.3 Tuberculosis among the Inuit

The Inuit are one of the 3 Canadian-born Aboriginal populations, along with the Métis and First Nations people. Approximately ¾ of Inuit reside in the Arctic, spanning Labrador to North West Territories (66). They represent the smallest proportion, at 4.2% compared to the Métis (32.3%) and the First Nations (60.8%), and comprise only 0.2% of the total population of Canada (66).

The first documented TB outbreak in the Canadian Inuit occurred in 1929 in Kugluktuk, in the territory now known as Nunavut (67). TB cases were subsequently reported in many Inuit communities, particularly those with the most frequent contact with Europeans (68). Before 1946, there was no systematic treatment of TB in the Canadian Inuit, thus the number of deaths attributed to TB in the North West Territories (NWT) was reported at 71.8 per 10,000 in 1950 (68). In the early 1950s, however, anti-tuberculosis treatment became available and aggressive TB control measures were implemented, with Inuit diagnosed with TB sent South to sanatoria for treatment. In addition, chest x-ray screening was initiated in some regions to identify active cases, and investigation of persons exposed to smear positive cases was conducted to identify individuals with prevalent disease (68). In 1968, collection of sputum samples from all persons with cough also became part of care (68). Over the course of the 1960s, rates of TB dropped in NWT from 109.5/10,000 to 17.7/10,000 (68). Despite this decline, in 1969, the rate of TB was still 50 times greater than the general Canadian population (69).

TB incidence in the Inuit plateaued during the 1980s (2). Since this time, however, it has steadily increased, such that in 2014, the incidence was 198.3 per 100,000 (64). Despite their small population size, the Inuit currently experience the highest burden of TB in Canada, approximately 330 times that of the Canadian-born non-Aboriginal population (64). A study in Nunavik, the Arctic region of Quebec from 1990-2000 suggested that the majority are due to ongoing transmission, as has been seen in other Aboriginal populations (2). It remains unclear why the rates of TB continue to be so drastically elevated in the Inuit, as there is no multi-drug resistance and minimal HIV co-infection. A better understanding of transmission dynamics and associated risk factors in this context is needed.

2.3 Approaches for delineating transmission

As previously discussed, TB disease can occur rapidly following initial infection, reflecting recent transmission, or years later in a process known as ‘reactivation’. A person can also be treated for TB and have a recurrence of the disease, either due to relapse or reinfection. The ability to discriminate between these events is crucial for TB control, as the population-level interventions differ substantially.

Clinical medicine predominantly relies on passive case finding, wherein patients self-present for evaluation and diagnosis. As the reasons for presenting are most often symptom-related, this approach can miss asymptomatic or minimally symptomatic individuals, who are still capable of transmitting to others. When most cases in a community are attributable to transmission, TB control programs should intensify efforts to find and treat these individuals. This involves a shift from conventional passive to active finding. Epidemiologic data can be used to identify clusters of transmission and associated risk factors, and therefore target public health interventions such as increased surveillance or screening to high-risk groups. In contrast to this, where the majority of cases in a community are attributable to reactivation of remote infection, TB control programs should focus on detecting persons with LTBI. Subsequent interventions should target acceptance of prophylaxis and treatment adherence, to prevent future instances of reactivation.

Recurrent TB can occur due to either relapse (of the same infection) or reinfection (with a new bacterium). The ability to distinguish relapse from reinfection is critical. Relapse after treatment suggests that therapy was inadequate, either due to inappropriate regime, adherence or drug malabsorbance, and requires intervention at potentially both patient and population levels. Recurrence of TB due to reinfection is more common in high-incidence settings, when there is ongoing transmission in the community, and hence, more opportunity for people previously treated to become re-infected.

In the following section, I will discuss different approaches that have been used to investigate transmission.

2.3.1 Contact investigation

In countries with low TB incidence such as Canada, contact investigations are routinely employed when a person is diagnosed with active TB. Such investigations aim to identify prevalent secondary cases as well as screen and treat individuals with *M. tuberculosis* infection (70, 71). Generally, a stone-in-pond principle is applied (72), wherein contacts are prioritized by proximity and likelihood of infection. If a contact is found to have TB infection or disease, it is typically attributed to the case under investigation. In the special circumstance wherein a

young child is first diagnosed with active TB, a 'reverse' contact investigation typically ensues, in order to identify the likely source of transmission.

There are several limitations to using this approach alone for delineating transmission. Firstly, it relies on the naming of contacts by the putative source case. Fear of social stigma, or contact associated with illicit behaviours may reduce the willingness of individuals to share such information (73-75), thereby limiting its efficacy to detect both secondary infections and disease. Secondly, because the investigation is based on proximity of contact to the putative source case, i.e., transmission from this source is presumed, additional potential sources of transmission (76) may be overlooked. This method is also unable to discriminate recent from remote infection. Infection can be detected by administering tuberculin skin tests during contact investigation. However, in absence of a recently documented negative test result, a positive result at this time could reflect transmission at any point during an individual's lifetime.

2.3.2 Molecular epidemiology

2.3.2.1 Methods

The idea of using repetitive insertion sequences (IS) in the *M. tuberculosis* genome to delineate transmission was first proposed in 1990 (77). IS*1081* and IS*6110* were shown to be present exclusively in species within the *M. tuberculosis* complex but the latter demonstrated high strain-to-strain variability, making this an ideal target for epidemiology (78).

Called 'restriction fragment length polymorphism' (RFLP), this molecular genotyping technique (described in (78)) involves extracting DNA from cultured *M. tuberculosis*. DNA is then digested using a restriction endonuclease (*PuvII),* which cleaves at the IS*6110* sequence. Gel electrophoresis is used to separate the DNA fragments by molecular weight. After transfer of these fragments to a membrane, those containing the IS*6110* element are identified using a DNA probe that tags the IS element on the right side of where it was cleaved. Chemiluminescense is then used to detect these probes, resulting in a banding pattern produced known as a 'DNA fingerprint'. Each patient's bacteria (referred to as an 'isolate') has its own fingerprint, facilitating comparison between the bacteria from one person and another.

In defined outbreaks, isolates from different patients were found to share the same DNA fingerprint (79). Conversely, randomly selected isolates from the same community had different DNA fingerprints. Thus, it has been inferred that when isolates from different patients share the same fingerprint (or have up to 1 difference, in some studies), this represents recent transmission. Conversely, when isolates present different fingerprints, transmission between the two individuals can be refuted. If no other isolates are identified with the same fingerprint in the same geographical location and within a reasonable timeframe, a case is attributed to reactivation of remote infection. Studies have shown the half-life of RFLP (a measure of the rate of change these patterns have in patients over time) approximately 2-3 years, supporting this interpretation (80, 81).

The number of IS*6110* bands present on RFLP is an important consideration. Strains with <6 bands ('low-copy') have poor discriminatory power, as there are fewer bands to compare between isolates. In this scenario, RFLP should be supplemented with a second molecular typing method (82). Such methods include mycobacterial interspersed repetitive units (MIRU) and spoligotyping (reviewed in detail in (82)). In brief, MIRU genotyping utilizes polymerase chain reaction (PCR) and pulse field gel electrophoresis to examine the size and number of repeated sequences in 24 different loci in the genome. The result is a 24-digit 'fingerprint' that can be utilized in similar fashion as RFLP for delineating transmission. The discriminatory power of MIRU is slightly lower compared to RFLP, except when low-copy strains are included (83, 84). Spoligotyping, an alternative method, exploits the presence of spacer sequences that separate repeats within the direct-repeat locus of the *M. tuberculosis* genome (82). By examining the presence or absence of specific spacers in the genome, strains can be compared. However, convergent evolution of these sequences (85) in unrelated strains can limit their utility in molecular epidemiologic studies, as matching patterns may not be indicative of recent transmission.

2.3.2.2 Important contributions to TB epidemiology

Molecular epidemiologic methods have made significant contributions to our understanding of TB transmission. Numerous studies have utilized these tools to investigate transmission, either to affirm epidemiologic links between patients or identify as yet unknown clusters of

transmission. For example, in a study of hospitals in New York, analysis of RFLP revealed that a high proportion of TB in health care workers was due to previously unsuspected occupational transmission (86). In another study in San Francisco, authors used RFLP to demonstrate that sputum smear negative cases, previously considered minimally contagious, were actually responsible for 17% of transmission (95% CI 12-24%) (40). Cross-contamination of patient samples in the diagnostic laboratory was also identified using RFLP, leading to the recommendation that possible laboratory contamination should be investigated in all persons with sputum smear negative results and single positive culture (87, 88).

At the population level, these methods have been used to evaluate trends in clustering over time (e.g., (89-91)). In San Francisco, Jasmer *et al.* found clustering decreased from 10.4 cases per 100,000 in 1991 to 3.8 per 100,000 in 1997 following initiation of additional TB control measures (89). Over the same period of time, the incidence of unique (non-clustered) cases was stable. This provided a natural control, as the TB control measures implemented would not be expected to influence rates of reactivation disease. In another population-based study, Borgdorff *et al.* (91) used RFLP to investigate the factors associated with declining TB rates in the Netherlands. Examining all cases diagnosed over 14 years, authors found that the decline was predominantly due to fewer instances of reactivation among those born in the Netherlands, in line with country-wide data on LTBI prevalence. Using the same dataset, authors also estimated the incubation period for progression from infection to disease among secondary cases. Out of 1095 epidemiologically-linked secondary cases who developed disease in the 15 years following exposure, 45% had progressed within the first year, 62% within 2 years and 83% within 5 years (92). This suggests that the estimated 10% risk of progression among those infected with *M. tuberculosis* is not evenly distributed over time. As risk appears to be highest in the first years immediately following infection, administration of LTBI prophylaxis is likely most warranted during this time.

In addition to delineating transmission networks and population-level, molecular methods have also contributed to our understanding of relapse versus reinfection. Population-based studies in low-incidence settings have demonstrated that most recurrent TB is due to relapse, rather than reinfection (e.g., (93)). It is not surprising that reinfection correlates with the prevalence of

active TB in a community, however other factors such as HIV also influence this risk. Using RFLP, a study in Malawi found that reinfection caused 12/23 recurrences of TB in HIV-positive patients compared to only 1/16 among HIV-negative patients (94). Without molecular genotyping techniques, distinguishing between these two events would be impossible.

Despite these important contributions, recent data have challenged the discriminatory ability of these tools. A newer method, whole genome sequencing, has shown resolution beyond that obtained with identical DNA fingerprints. Discussed in detail in the following chapter, the rise of WGS has led to a paradigm shift in the field of molecular epidemiology.

CHAPTER 3. REVIEW OF THE LITERATURE PART 2 – WHOLE GENOME SEQUENCING AND GENOMIC EPIDEMIOLOGY OF TB

3.1 The new era of "genomic" epidemiology

The first complete bacterial genome, *Haemophilus influenzae*, was sequenced in 1995 over a period of several months using a method called 'Sanger sequencing' (95). The estimated cost of sequencing the 1.8 Mega-base genome was 48 cents per base-pair, or ~$864,000. Since this time, considerable scientific advances have been made in sequencing technology (96, 97). While classical Sanger-based methods relied on first synthesizing and then detecting the DNA sequence, 'next-generation sequencing' (NGS) performs both tasks simultaneously. The development of high-throughout approaches has provided the capacity to sequence large numbers of genomes within days, and while was ~120 GBP (12) per genome in 2015, this cost continues to decline. This sequencing revolution has fuelled the use of WGS in what has come to be termed 'genomic epidemiology'.

3.2 An introduction to whole genome sequencing

Classical molecular typing methods interrogate ~1% of the *M. tuberculosis* genome. In contrast, whole genome sequencing allows the interrogation of all 4.4 million base-pairs. The following is a brief introduction to this genotyping tool, to facilitate understanding of the epidemiologic applications described in this chapter. Please see **Chapter 4** for a detailed description of WGS methodology and techniques used throughout this thesis.

Notwithstanding the fact that recent attempts have been made to obtain *M. tuberculosis* sequence reads from raw clinical samples (98, 99), (98, 99), to obtain high-quality WGS data for epidemiologic purposes, genomic DNA must currently be extracted from pure cultures of *M. tuberculosis*. Once extracted, this DNA is fragmented into segments of a desired size, which are amplified, and then sequenced by synthesis. The sequences generated from these fragments of DNA are called 'reads'. Long-read sequencing platforms (such as PacBio) can produce reads that are >10 kilobases in length and optimally suited for completing genomes, in a process called '*de novo* assembly.' This entails comparing each read with one another, and using overlapping segments to determine the complete genomic sequence. Short-read sequencing

platforms, such as the Illumina MiSeq, can produce reads that are up to 300 bp in length. Such reads are typically aligned ('mapped') to a pre-existing, complete reference genome to re-build the genome under investigation. Short-reads tend to have lower error rates per nucleotide than long read sequencers (97), but are not optimal for long, repetitive sequences; accurate mapping is difficult because short-reads do not span such regions. Therefore, these regions are frequently excluded from analysis (100, 101).

The ability to perform a reference-based analysis is dependent on availability of high-quality, complete genomes for alignment. Because some such genomes are available for *M. tuberculosis* (e.g., lineage 4 references H37Rv and CDC1551), most genomic epidemiology studies have utilized short-read data and aligned results to one of these references genomes. The consequences of using a potentially divergent reference, i.e., from a different lineage of *M. tuberculosis*, have not previously been investigated and are therefore examined in **Chapter 8.** As many such aspects of data analysis have yet to be validated, this represents an important step needed for bioinformatics pipeline standardization.

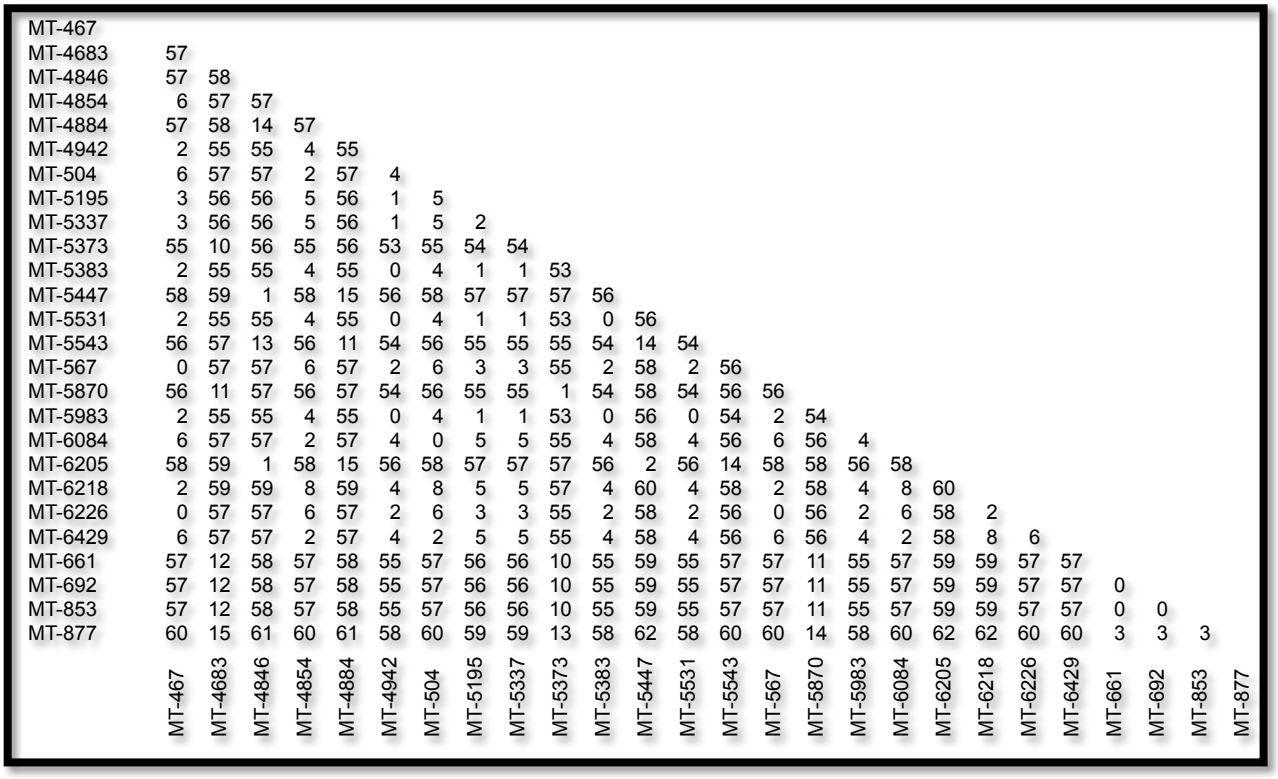
The precise steps and tools utilized in current bioinformatics pipelines vary widely between studies. However, overall most WGS analyses will perform some degree of quality control both before and after the initial alignment. These may include removing PCR duplicates retained from the library preparation stage and locally re-aligning around insertions and deletions. Single nucleotide polymorphisms (SNPs, differences in a single base) are then identified ('called') compared to the reference genome. After this, SNPs are then annotated to determine the location and functional implications of the mutations identified. These SNPs are then filtered for quality to reduce the number of false positives due to mapping or sequencing error, or to exclude SNPs from certain genes, as warranted. This final SNP dataset is then used to compare different *M. tuberculosis* genomes for epidemiologic purposes.

3.2.1 Key outputs of whole genome sequencing for epidemiology

3.2.1.1 SNP matrices

SNP matrices provide the pairwise distances, measured in single nucleotide polymorphisms, between isolates. These may be used to help resolve transmission, discussed further in section

3.5. As shown below in **Figure 3-1**, some isolates may have zero or few SNPs between them, while other may have many. The implications of this are discussed below in section 3.2.3.



	MT-467	MT-4683	MT-4846	MT-4854	MT-4884	MT-4942	MT-504	MT-5195	MT-5337	MT-5373	MT-5383	MT-5447	MT-5531	MT-5543	MT-567	MT-5870	MT-5983	MT-6084	MT-6205	MT-6218	MT-6226	MT-6429	MT-661	MT-692	MT-853	MT-877
MT-467																										
MT-4683	57																									
MT-4846	57	58																								
MT-4854	6	57	57																							
MT-4884	57	58	14	57																						
MT-4942	2	55	55	4	55																					
MT-504	6	57	57	2	57	4																				
MT-5195	3	56	56	5	56	1	5																			
MT-5337	3	56	56	5	56	1	5	2																		
MT-5373	55	10	56	55	56	53	55	54	54																	
MT-5383	2	55	55	4	55	0	4	1	1	53																
MT-5447	58	59	1	58	15	56	58	57	57	57	56															
MT-5531	2	55	55	4	55	0	4	1	1	53	0	56														
MT-5543	56	57	13	56	11	54	56	55	55	55	54	14	54													
MT-567	0	57	57	6	57	2	6	3	3	55	2	58	2	56												
MT-5870	56	11	57	56	57	54	56	55	55	1	54	58	54	56	56											
MT-5983	2	55	55	4	55	0	4	1	1	53	0	56	0	54	2	54										
MT-6084	6	57	57	2	57	4	0	5	5	55	4	58	4	56	6	56	4									
MT-6205	58	59	1	58	15	56	58	57	57	57	56	2	56	14	58	58	56	58								
MT-6218	2	59	59	8	59	4	8	5	5	57	4	60	4	58	2	58	4	8	60							
MT-6226	0	57	57	6	57	2	6	3	3	55	2	58	2	56	0	56	2	6	58	2						
MT-6429	6	57	57	2	57	4	2	5	5	55	4	58	4	56	6	56	4	2	58	8	6					
MT-661	57	12	58	57	58	55	57	56	56	10	55	59	55	57	57	11	55	57	59	59	57	57				
MT-692	57	12	58	57	58	55	57	56	56	10	55	59	55	57	57	11	55	57	59	59	57	57	0			
MT-853	57	12	58	57	58	55	57	56	56	10	55	59	55	57	57	11	55	57	59	59	57	57	0	0		
MT-877	60	15	61	60	61	58	60	59	59	13	58	62	58	60	60	14	58	60	62	62	60	60	3	3	3	

FIGURE 3-1. Example of a pairwise SNP matrix.

3.2.1.2 Phylogenetic trees

Phylogenetic trees, also known as 'dendograms' are visual representations of the relatedness between bacterial isolates. These can be produced using data from classical molecular typing methods, as well as WGS. In the latter approach, SNPs differentiating each bacterial isolate from the reference genome are used to produce these trees via a variety of methods. Closely-related isolates, which share unique SNPs that are not found in any other isolates, are grouped together in 'clusters'. Phylogenetic trees can be used to help rule out transmission, as patients with isolates in different clusters are very unlikely to have transmitted to one another. In contrast, transmission from person to person is more likely to have occurred within clusters.

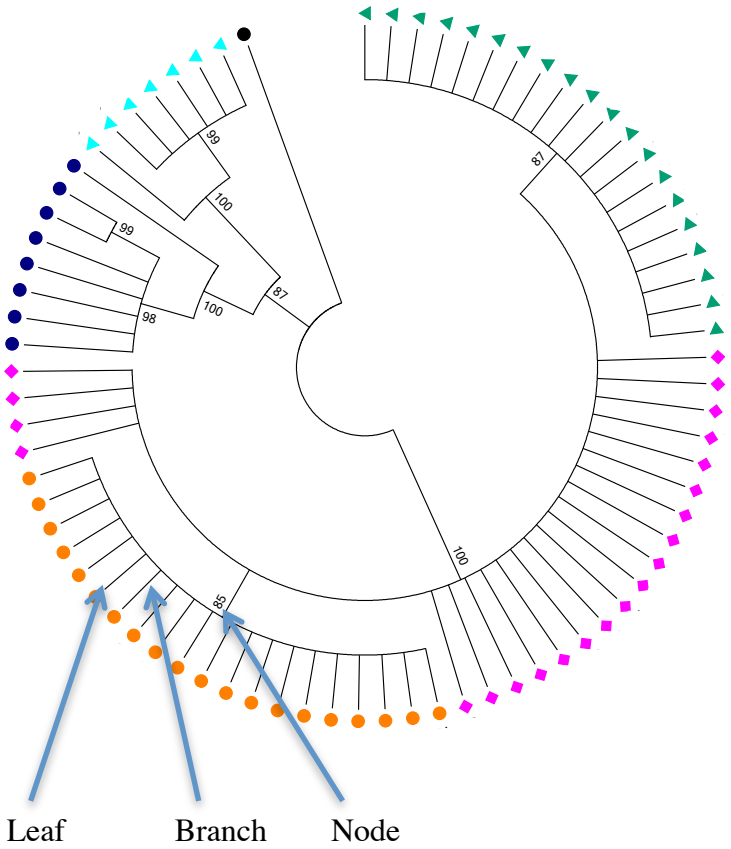


FIGURE 3-2 Example of a phylogenetic tree. Distinct clusters are indicated, with all isolates from a cluster depicted in the same colour. Branches with bootstrap proportions under 80% (102), have been collapsed as these indicate low accuracy. As such, branch lengths do not correspond to absolute genetic distance. However, it can be interpreted that isolates ('leaves' or 'tips') that belong to the same cluster are most closely related to one another, as these share the same most recent node. Moving proximally in the tree from the orange tips, the first node, therefore, indicates a missing recent common ancestor of these orange isolates. The second proximal node is shared with the isolates in pink. This indicates a common ancestor between the orange and pink, which occurred at some time further in the past.

3.3 Comparing WGS to classical molecular typing methods for tuberculosis

As discussed, using classical molecular typing methods, recent transmission was a dichotomous event; if two patterns on RFLP, MIRU or spoligotyping were the same or highly similar, this was suggestive of recent transmission between the pair, while different patterns indicated reactivation of remote infection. In 2009, a study by Niemann *et al.* (103) challenged this interpretation. Applying WGS to two RFLP-identical isolates with differing drug susceptibility profiles, authors revealed that the pair were differentiated by 130 SNPs and a large deletion, and therefore definitively refuted that they were part of the same transmission network.

Shortly afterwards, WGS was applied to a complex TB outbreak in the Netherlands (104). All 104 isolates within this outbreak, spanning over 18 years, shared identical RFLP fingerprints. Contact investigation only detected two probable transmission chains, with directionality of transmission between individuals and other potential networks remaining largely unresolved. By performing WGS on 3 outbreak isolates, authors were able to identify 8 SNPs that could be used to discriminate these isolates. Testing the remaining 101 isolates from the outbreak for these 8 SNPs revealed 5 distinct clusters of transmission within the single homogenous group previously identified by RFLP, and combined with clinical epidemiologic data, increased understanding of chains of transmission.

The following year, in 2011, WGS was utilized for the first time to describe an entire outbreak in British Columbia (3), with 32/37 outbreak isolates sequenced as well as 4 historical genomes from the same region. Whereas MIRU identified one outbreak with one probable source case, WGS revealed two distinct groups of transmission that had been circulating in the community for at least 5 years. This study was the initial impetus for using WGS in analysis of the Northern outbreak. Since this time, the higher resolution of WGS in comparison to classical methods has since been further substantiated by a number of studies (101, 105-109).

3.4 Using WGS to determine transmission

In addition to more detailed resolution of clusters, and in contrast with classical molecular tying methods which can only suggest or refute transmission, WGS may be used to infer directionality of such events. New mutations in the *M. tuberculosis* genome are acquired over time, and are not thought to revert back to wild-type (110). As *M. tuberculosis* is thought to be highly clonal, with minimal horizontal gene transfer (111, 112) or homologous recombination outside the repetitive proline-glutamate (PE) and proline-proline-glutamate (PPE) genes (113), one can follow the acquisition of SNPs as TB is transmitted from person to person, potentially providing greater accuracy in delineating chains of transmission. The maximum number of SNPs that can occur during such transmission, however, may depend on a number of factors.

3.4.1 The molecular clock of M. tuberculosis and SNP thresholds

An important consideration in determining transmission via WGS is the rate at which mutations occur over time, i.e., the 'molecular clock' of *M. tuberculosis.* Studies in macaques have suggested a rate of ~0.39 SNPs per genome per year for active TB (95% CI 0.16-0.80), with similar results for during latent infection in these same experimentally infected animals (53) Using pairs of human isolates collected longitudinally from the same patients as well as across patients in household outbreaks in the UK, Walker *et al.* estimated the mutation rate of *M. tuberculosis* to be 0.5 SNPs per genome per year (95% CI 0.3-0.7) (105) in accordance with the study by Ford *et al.* (53). In contrast, during latent infection, a small study involving 4 pairs of patients by Colangeli *et al.* suggested that the mutation rate in humans decades after exposure may be one log slower than in the two years preceding active disease (114).

Based on these observations, authors have proposed thresholds for how many SNPs can separate cases that are linked by transmission. For instance, Walker *et al.* classified isolates from cases separated by 12 or more SNPs as unlikely to be transmission, called those with 6-12 SNPs as indeterminate and classified isolates separated by 5 or less SNPs as probable transmission. Combining household outbreaks and community-based MIRU clusters, all 69 pairs of cases with epidemiologic links suggestive of transmission had isolates within 5 SNPs; however, 62 of 75 pairs (83%) without epidemiological links also had isolates under this threshold making the interpretation of this result unclear.

A similar mutation rate (0.4 SNPs per genome per year) was identified by Roezter *et al.*, wherein authors sequenced 86 isolates from a single RFLP-defined outbreak in Germany. Classical typing methods suggested a single source, in contrast to both the spatial distribution of cases and contact investigation data. Using WGS, authors resolved the 'outbreak' into 7 distinct clusters of 2 to 24 isolates. The majority of these clusters were closely related (called the 'Hamburg clone'). By examining the accumulation of SNPs over time (and deletions), transmission events leading to introduction and subsequent dissemination of this clone in two other cities were identified. The resolved clusters also agreed with contact investigation data, where epidemiological links between 31 patients were resolved in 8 different transmission

chains. Within these pairs of patients, a maximum of 3 SNPs distance was seen, however no time period between these pairs was provided.

Another study in San Francisco, while not directly estimating the molecular clock, examined transmission in a 22-month outbreak involving 9 cases (115). WGS revealed 0-2 SNPs between all direct person-to-person transmissions, and facilitated the identification of a previously undiscovered epidemiologic link between two cases of probable transmission.

Bryant *et al.* inferred a slightly slower rate, at ~0.3 SNPs per genome per year, using isolates from 199 patients in Amsterdam. However, this rate varied substantially by method used and between isolates. 185 of these patients had known epidemiologic links and comprised 42 RFLP-clusters, while the remaining 14 were from the same clusters, but had no known links. Those with epidemiologic links had isolates separated by a median of 2 SNPs (range 0-149 SNPs). As in the UK study (105), however, low numbers of SNPs were frequently found between pairs lacking epidemiologic support. 82 of such pairs, separated by a maximum of ~5 years, had 0 SNPs between them. While undetected transmission is a possibility, without epidemiologic confirmation, the WGS data alone was insufficient to 'rule in' transmission. In contrast, authors were able to use WGS to effectively 'rule out' such events; one pair of RFLP-identical isolates from patients with a known epidemiologic link were 149 SNPs apart as well as 2 independent deletions, refuting transmission.

3.4.2 Within-host diversity: micro-evolution and/or mixed infection

Another variable that may complicate our interpretation of transmission is the potential for within-host diversity. Small mutations may occur within a host as *M. tuberculosis* replicates over time, such that some bacteria have the original allele while others in the same host have acquired a SNP at that locus. This is called 'micro-evolution'. Perez-Lago *et al*. (116) found that repeat respiratory specimens obtained within +/- 1 day from the same patients can differ by as many as 7 SNPs. SNPs were also found between respiratory and non-respiratory specimens from the same patients, albeit to a lesser degree. However, as this study was based on samples collected from only 4 patients, it may not be truly representative of within-host diversity across different sites. Additionally, while authors suggest this diversity is due to micro-evolution

within the host, patients were pre-selected based on MIRU variability between samples. Depending on the local strain diversity and prevalence of TB, an alternative explanation for such variation might be infection from >1 source ('mixed infection'), as identified in (117).

Within-patient diversity was also examined in Walker *et al.* (105), using 49 pairs of cross-sectional isolates from respiratory and non-respiratory sites collected within the same month. In this study, 79% of pairs were 0 SNPs apart, while 96% of pairs (47/49) were <5 SNPs. For the remaining 4 isolates, one pair was separated by 11 SNPs, while the other was separated by >400 SNPs, indicative of either mixed infection or laboratory contamination. No cross-sectional comparison was made between respiratory samples from the same patients. Paired longitudinal respiratory samples were available from 30 patients, with 28 of these presenting 0-10 SNPs difference. However, such samples were taken 6 to 102 months apart with no time-stratified data provided.

Transmission does not typically occur from extra-pulmonary sites (excepting during invasive medical procedures). The diversity between respiratory samples from the same host, rather than between respiratory and non-respiratory sites, is therefore of primary interest as this can potentially be transmitted forward. Regardless of the cause, both micro-evolution or mixed infection may obscure transmission, influencing epidemiologic inferences. Methods to detect such diversity are therefore discussed in **Chapter 4**.

3.4.3 Sampling

Molecular (or genomic) epidemiologic studies are typically limited by the number of isolates available from cases, as well as the time period and geographical location under study. While no studies have been published examining the effects of sampling using WGS data, previous work conducted using RFLP indicated that the sampling fraction, (i.e., the proportion of total cases included in the study) has the greatest influence on clustering, with a higher proportion of clusters identified as sampling fraction increases (118, 119). Identifying a cluster of transmission requires that at least 2 isolates from the same group are included in the study. The impact of sampling fraction on the proportion of cases clustered and estimates of recent, ongoing transmission are greater when the true underlying size of the clusters in a community is

small (118). In this scenario, as lower fractions of cases are sampled, a higher proportion of 'true' clustered cases are likely to be misclassified as 'unique'. Another study, using computer simulations to model transmission under settings with varying degrees of TB transmission and cluster size, suggested that as sampling fractions decreased, the odds ratios for risk factors associated with clustering were systematically biased towards the null (120). Conversely, using real data from >12,000 TB cases over 15 years in the Netherlands, Borgdorff *et al*. found that risk factors for clustering were similar in magnitude - albeit less precise – when randomly sampling as low as 40% of cases. Authors also compared odds ratios for clustering when restricting by study duration (1, 2, 4 or 8 years), or geographical boundaries (the whole country vs. by postal code or infectious disease control unit) (119), with generally stable results; because risk factors for clustering were also risk factors for belonging to a large cluster, this minimized the impact of misclassification of isolates from smaller clusters that were undetected at lower sampling fractions. Given these analyses are based on a low-incidence setting, in an environment with higher rates of clustering and transmission, such stability of the odds ratios may not occur.

3.4.4 Laboratory cross-contamination

As with classical molecular typing methods, it is important to rule out laboratory cross-contamination. This is an issue in mycobacteriology labs, as a number of patient specimens are processed together (called 'batched processing'), with the resultant risk that a single positive sample can inadvertently spill into cultures for other specimens if there is a break in laboratory technique. In the WGS era, isolate pairs separated by zero SNPs should be investigated to determine if these were processed the same date, as part of the same clinical workflow in the same laboratory. Potential cross-contamination should be suspected particularly for patients with smear negative disease and a single positive culture.

3.4.5 Local strain diversity

Local strain diversity may be an important consideration in determining whether cases are due to transmission or reactivation. Analysing *M. tuberculosis* from patients in Arkansas in 1992-1993, authors found that 78 (33%) of cases with RFLP and secondary probe, pTBN12, shared

the same fingerprint with at least one other patient but only 42% of secondary cases had epidemiologic links with a member of the same cluster (121). Of those without links, the majority (64%) had never resided in the same counties as the other members of their clusters. They were also older than those with epidemiologic links (mean age 58 +/-16 versus 45 +/- 14) and 31% had a history of previous infection or exposure to TB decades in the past. This suggested to public health that their current TB episode was a result of remote activation of infection that had occurred decades prior, rather than ongoing transmission. A similar case-control study in Québec further highlights the importance of local strain diversity in interpretation of molecular (or genomic) epidemiologic data (122). Compared to controls, all 77 cases shared a unique mutation in the *pncA* gene, conferring resistance to pyrazinamide. These cases had similar fingerprints on RFLP, and identical spoligotype patterns. As there were no epidemiologic differences between cases and controls, the latter representing a diversity of genotypes and lacking the *pncA* mutation, the appearance of this drug-resistant strain was attributed to reactivation of a previously endemic strain rather than ongoing transmission.

3.5 WGS to differentiate relapse from reinfection

Two studies have utilized WGS to differentiate relapse from reinfection (117, 123). As part of the REMoxTB randomized controlled trial, Bryant *et al.* examined pairs of isolates from 47 patients diagnosed with recurrent TB in Malaysia, South Africa and Thailand (117). Excluding 5 pairs wherein the second culture was likely laboratory contamination, 33 pairs were identified as 'relapse' by WGS. All had fewer than 7 SNPs between them. Surprisingly, 6 of these pairs were classified as 'reinfection' by MIRU – one with 4 MIRU loci difference despite having zero SNPs between isolates. This suggests that misclassification is possible between MIRU-defined clusters, as well as within. In terms of reinfection, 3 pairs of isolates were classified thus by both typing methods, with >1300 SNPs and at least 3 MIRU loci difference between them. The remaining 6 pairs were determined to be mixed infection, 4 of which were undetected by MIRU.

Similar results were found in a study by Guerra-Assuncao *et al*. in Malawi (123). Considering only pairs of isolates from different episodes of TB (i.e., recurrence), there was a clear divide between relapse and reinfection. All pairs classified by WGS as 'relapse' were under 9 SNPs

while those classified as reinfection were all separated by >100 SNPs. Including pairs of isolates from the same episode of TB, all under 9 SNPs had identical RFLP or a single band difference, while all with >100 SNPs had more than one band difference on RFLP. No mixed infection was identified.

Both studies show a clear distinction between relapse and reinfection with WGS, with two very different SNP distributions for each event. While Bryant *et al.* suggest this may be due to immunological protection against reinfection by closely-related strains, it is important to note that both studies occurred in high-incidence settings, with substantial strain diversity. An alternative explanation is that the probability of being re-infected is greatest with a genetically-dissimilar strain simply because these are widely circulating in the community.

3.6 Application to surveillance

In addition to retrospective analyses, several studies have suggested a role for WGS in ongoing, real-time TB surveillance. However, as WGS is still more expensive than conventional methods, cost is a potential obstacle for many local public health departments. The technical expertise both to conduct and analyse WGS data is also limiting, and unlikely to be decentralized in the near future. In attempt to alleviate these issues, Stucki *et al.* developed a SNP-based assay for a local outbreak strain, based on WGS of 3 historical outbreak isolates. Using this assay, they were able to rapidly screen 1,642 isolates, and identify 68 ‘outbreak’ isolates for further analysis and resolution by WGS. A similar study was conducted in Spain, using a PCR-based test to identify predominant strains circulating in the community (110). While such tests have the ability to rapidly identify known strains and be performed easily in local laboratories rather than outsourcing to reference laboratories, it is important to note that novel strains would not be captured using this approach. Therefore, periodic monitoring in each region via WGS and subsequent updating of the genomic targets of such assays to ‘capture’ new, emerging strains would be warranted. Furthermore, as variation between isolates would only be detected in the targeted SNPs (‘phylogenetic discovery bias’), the level of discrimination and utility of such a tool in epidemiology would be greatly influenced by the precise SNPs and number selected.

3.7 WGS and the evolution of drug resistance

WGS has also provided valuable insights into the development and spread of drug resistant tuberculosis at the population level. WGS studies have illustrated that drug resistance can occur in a step-wise fashion over the course of decades and once developed, spread throughout communities (124, 125). In the case of an XDR outbreak that was first documented in KwaZulu-Natal in 2005 (125), RFLP and spoligotyping indicated that a single predominant drug-resistant strain was transmitting through the community. Using historical isolates as well as strains closely related to the XDR strain, authors confirmed the clonal nature of this outbreak and examined the sequence at which mutations were acquired in the community over time. WGS revealed that resistance to isoniazid and streptomycin was first acquired 50 years earlier, with subsequent mutations conferring resistance to ethambutol and ethionamide, followed by rifampin and then pyrazinamide. Resistance to second-line drugs in occurred in the 1990s, ultimately culminating in the same resistance pattern that would characterize the future XDR outbreak in 1995. Importantly, acquisition of resistance to each drug was in line with its discovery and introduction into clinical management, and not driven predominantly by current TB control or HIV in this population. When authors examined strains that were not part of the outbreak, they also observed numerous instances of independent acquisition of multi-drug resistance and evolution to pre-XDR status. Resistance consistently commenced with isoniazid, which is not targeted by GeneXpert, the current molecular test utilized in this region (125).

Investigation of an outbreak of MDR-TB in Argentina revealed similar findings (124). While the outbreak was first documented in the early 1990s and continues today, resistance to isoniazid was present in all isolates sequenced, and therefore was present in the most recent common ancestor, in 1970 (95% CI 1966-1975). Resistance to rifampin and streptomycin were acquired in 1973 (95% CI 1968-1978), while additional resistance to pyrazinamide, ethambutol and kanamycin was estimated to have emerged in 1979. For most drugs, emergence of resistance was in sequence with wide-spread use in TB treatment.

Mutations conferring drug resistance often come at a cost to bacterial fitness; the extent of the effect on fitness determines how frequently such mutations are seen in clinical isolates, i.e., how transmissible they are (126). Compensatory mutations, which reduce this fitness cost, can

drive epidemiologic success of a bacterium. Using WGS, Casali *et al.* explored the epidemiology of drug resistance and role of such mutations in Russia (127). Of 1,000 prospectively collected isolates from patients, 64% were closely related, comprising a single Beijing sublineage with two predominant clades. As older populations of bacteria are more genetically diverse, this suggested a recent introduction and subsequent expansion via transmission of this sublineage in the region. High amounts of resistance were identified in these isolates, with 74% and 70% having mutations in *katG* and *rpoB*, conferring resistance to isoniazid and rifampin, respectively. Authors demonstrated that compensatory mutations in genes *rpoB* and *rpoC*, among others, were associated with the *rpoB* mutation conferring resistance to rifampin, potentially contributing to the success of MDR strains in this population (127). This is in contrast to the Argentinian outbreak (124), wherein no association between *rpoC* and strain fitness was identified, as only 2/21 such mutations were identified in related isolates.

3.8 Summary

When this thesis work was initiated in 2012, the utility of WGS to TB epidemiology was still unclear. However, as this tool became more feasible and widely utilized, the increased information provided by WGS in comparison to older, classical molecular typing methods was evident. Using this method, we are now more accurately able to resolve transmission networks, unequivocally identifying source cases (105) within clusters of transmission, and even revealing missing (unsampled) cases that had transmitted to others (105, 116). WGS has also demonstrated increased ability to discern recent transmission from reactivation, reinfection from relapse and identify mixed infection. Finally, WGS has taken on a pivotal role in our understanding of the epidemiology of drug resistance. As such, many now consider this the 'gold standard' in molecular epidemiology. However, wide-spread use of this tool outside the research domain is still lacking in part due to the complex nature of the analysis. As outlined in the following chapter, substantial bioinformatics expertise is required to transform raw genetic data into an epidemiologic tool and as yet, this analytic pathway has not been standardized.

CHAPTER 4. MATERIALS AND METHODS

In this chapter, I have provided an overview of methodology used in each manuscript. As methods are also presented in the main manuscripts (for I-IV), and supplementary material (for I-III), I have aimed to build on what has already been described.

4.1 Study populations and data sources

The first three Objectives focus on the epidemiology of TB among the Inuit of Nunavik. Nunavik is the Arctic region of Quebec. It spans 443,685 km² and is comprised of 14 Inuit communities. The total population is 12,090 (128) as of 2011, with little out-migration from communities. Each is separated from the nearest village by a median distance of 137 km (interquartile range 110-178), with no connecting roads.

In Québec, TB is classified as a "Maladie à déclaration obligatoire", meaning physicians are required to report cases to their local public health department. The Nunavik Regional Board of Health and Social Services (NRBHSS) is the public health unit responsible for monitoring TB cases in Nunavik, and has collaborated on these projects. In addition, all specimens from TB suspects in Nunavik are processed at the mycobacteriology laboratory of the McGill University Health Centre (MUHC), ensuring 100% case ascertainment for this remote Artic region.

Objective 4 utilized these data to explore methodological issues related to WGS, while Objective 5 is a narrative review.

4.1.1 Genetic data

All TB samples used in this thesis work were provided by the Mycobacteriology Laboratory of the McGill University Health Centre and the Laboratoire de Santé Publique du Québec.

4.1.2 Clinical epidemiologic data

While assisting with the public health response during the outbreak, I developed a database to facilitate clinical follow-up of cases and their contacts in this village. When individuals were diagnosed with active TB, they were asked to provide detailed lists of all household and non-

household contacts. Contacts were then interviewed by trained health care professionals, using standardized data collection tools. Data was collected in real-time, including the results of these contact investigations, diagnostic test results, demographic data and information on risk factors such as cigarette smoking. Cavity on chest x-ray was updated retrospectively, based on a blinded clinical review by two independent respirologists after the 'outbreak'. Discordance was resolved using radiology reports from time of diagnosis. Sputum smear and culture results were also validated retrospectively with the Mycobacteriology Laboratory of the MUHC. This dataset was used with the permission of the NRBHSS for this research, as part of an ongoing collaboration with this unit.

For the years preceding the outbreak in this village and other villages of Nunavik, contact investigation data were validated and provided by Dr. Jean-Francois Proulx, of the NRBHSS. Age, sex, dates of diagnosis, sputum smear and culture results were provided by the Mycobacteriology Laboratory of the MUHC. Village at diagnosis was provided by the NRBHSS.

4.1.3 Data linkage and ethics

I performed linkage of genetic and epidemiologic data in nominal form, working under a professional mandate from the NRBHSS. All projects were conducted with ethics approval from the McGill University Faculty of Medicine's Internal Review Board, in collaboration with the village council and the NRBHSS. Individual patient consent was not required.

TABLE 4-1. Overview of sampling frame, by Objective

Objective	Study design	Sampling frame	Time Period	Total number of confirmed cases	Total number of cases with WGS successfully completed	Total number of pairwise SNP comparisons
1	Case-series	Villages of Nunavik, excluding the outbreak village	2006-2012	45	42 (93%) [a]	631 'improbable' transmission pairs (no epidemiologic links, from different villages)
	Case-series	Cases from outbreak village	1990-2010	32	29 (91%) [b]	3,003
			2011-2012	50	49 (98%) [c]	
2	Case-series	Cases from all villages of Nunavik	1990-2000	51	26 (51%) [d]	13,203
			2001-2013	149	137 (92%)	
3	Case-control	All persons with new infection in the outbreak village	2011-2012	34	34 (100%) [e]	[redacted]
4	Case-series	Same as Objective 2	[redacted]			
5	Narrative review	Medline/PubMed	1946 – Aug. 2015	[redacted]		

[a] 3 isolates not available for sequencing [b] 2 isolates not available for sequencing, 1 WGS low-quality. [c] WGS low-quality. [d] Convenience sample [e] Subset of cases from the outbreak village used in Objective 1

4.2 Inclusion and exclusion criteria

Objectives 1, 2 and 4:

As this work involved sequencing of bacterial DNA to track transmission, these studies have only included cases with at least one culture positive for *M. tuberculosis*. Therefore persons with clinical diagnoses of TB, without microbiologic confirmation, were not eligible.

Cases were classified as ‘pulmonary TB’ if at least one culture from sputum, lung biopsy or bronchoalveolar lavage grew *M. tuberculosis.* All cases during the ‘outbreak’ had exclusively pulmonary disease.

Objective 3:

This study aimed to investigate the association between exposure to active TB (to any case or to different genotypes, as identified in Objective I) and the progression to disease after recent infection. Therefore, the study population was restricted individuals with documented ‘new’ infection. New infection was defined by TST conversion or a documented new positive TST without any previous testing. Cases were defined as those with micro-biologically confirmed active TB, while controls were those who did not progress to disease.

Objective 5:

A narrative review was performed to investigate the potential role of WGS in the clinical diagnosis workflow for TB. Medline (1946-present, including in-process and non-indexed) and PubMed (for articles exclusive to PubMed) were searched on Aug. 6, 2015 using the terms “whole genome sequencing” and “tuberculosis” combined with “epidemiology” or “clinical”. No language restrictions were applied. Original studies describing the use of WGS in either TB diagnostics or prediction of drug resistance were included. Abstracts, conference proceedings, letters and review articles were not eligible. Reference lists of included articles and relevant review articles were also hand-searched for additional manuscripts.

4.3 Workflow for genetic data

4.3.1 Collection of M. tuberculosis samples

In Nunavik, all individuals with clinical suspicion of active TB are asked to provide 3 spontaneous sputum samples for smear microscopy and culture. During the ‘outbreak’, those who were unable to provide spontaneous samples underwent nurse-supervised sputum induction using 3% hypertonic saline. Pediatric patients were transported to the nearest hospital for gastric aspirate until May 2012. After this time, sputum induction with hypertonic saline and nasopharyngeal suctioning was performed at the local CLSC. This method is considered an acceptable alternative to gastric aspiration (6) and in some studies has a higher yield for pediatric

TB than gastric aspirates (129). Patient samples are transported by plane via Kuujjuaq to Montreal, where they are taken to the Biosafety Level 3 facility of the MUHC. Samples are stored at 4^{o}C until processing.

4.3.2 Sample processing

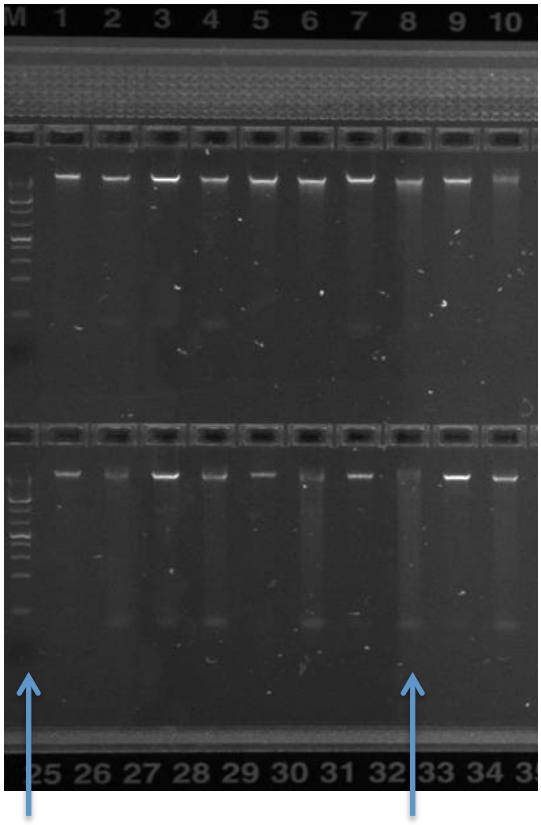
Patient samples were received, they were decontaminated used N-acetyl-L-cysteine and 2% sodium hydroxide in order to prevent over-growth by faster growing micro-organisms (130). After 20 minutes of treatment, samples were centrifuged and the sediment was re-suspended for inoculation into culture media.

4.3.3 DNA extraction

Mycobacterium tuberculosis was grown in culture using the BACTEC™ MGIT™ 960 Mycobacterial Detection System (Becton, Dickinson and Company). Isolates were then sub-cultured once on Middlebrook 7H10 agar supplemented with Oleic acid Albumin Dextrose Catalase enrichment media (Becton, Dickinson and Company) and DNA extractions were performed as per (131). All colonies on the plate were included, i.e., a full sweep of the plate was done.

4.3.4 DNA quantification

Genomic DNA (gDNA) quantity was first assessed by the McGill University and Génome Québec Innovation Centre using a fluorescence assay (the PicoGreen® double-stranded DNA Assay Kit from Life Technologies). 1-5 µg of gDNA are recommended by Illumina for sequencing. Gel electrophoresis was used to evaluate the condition of the DNA, i.e., to check for degradation or RNA contamination.



Ladder DNA degradation

FIGURE 4-1. Gel electrophoresis for assessment of DNA quality after extraction. Each lane represents DNA extracted from a different patient isolate. A 1 kiloBase DNA ladder was run. An example of DNA degradation is indicated, wherein all DNA was broken into lower molecular weight fragments. Potential RNA contamination is visible at the bottom of some lanes (<75 bp according to the ladder).

4.3.5 DNA library preparation

Library preparation and sequencing were also performed by the McGill University and Génome Québec Innovation Centre. All WGS was done using the Illumina MiSeq platform with 250 base-pair reads were generated. Sequencing was conducted from both ends of the DNA fragments ('paired-end'), using the protocol described below.

DNA 'libraries' were prepared for each sample using the TruSeq DNA High-throughput protocol (132). First, gDNA was fragmented by sonication into pieces 800 base-pairs (bp) or fewer in length (133). This process creates overhanging 5' and 3' ends of gDNA, which must then be blunted. A DNA polymerase adds bases to the former, while an exonuclease cleaves bases from the latter (133). To prevent self-ligation, an 'A' was then added to 3' end of these fragments

(133). Short adaptor sequences were then ligated to both the 3' and 5' ends of the gDNA fragments. For our analyses, TruSeq v1 adaptors were used.

After adaptor ligation, the DNA fragments were then purified by gel electrophoresis, with fragments correctly ligated to adaptors selected based on size. For paired-end sequencing with a desired read length of 250 bp (denoted as '2x250 bp'), Illumina suggests targeting fragments of 300 bp or greater (133).

After size selection, PCR was then used to amplify these DNA fragments (133). This step was used to add a final sequence to the adaptors that facilitates binding to the flowcell, i.e., the solid surface on which sequencing is performed, essentially anchoring the DNA fragments in place. PCR amplifies the fragments that have adaptors on both ends, necessary for such binding, removes any remaining adaptors that have self-ligated and by amplification, provides sufficient material for library quantification in the next step (133).

Such quantification was done by fluorescence using the Agilent Technologies 2100 BioAnalyzer.

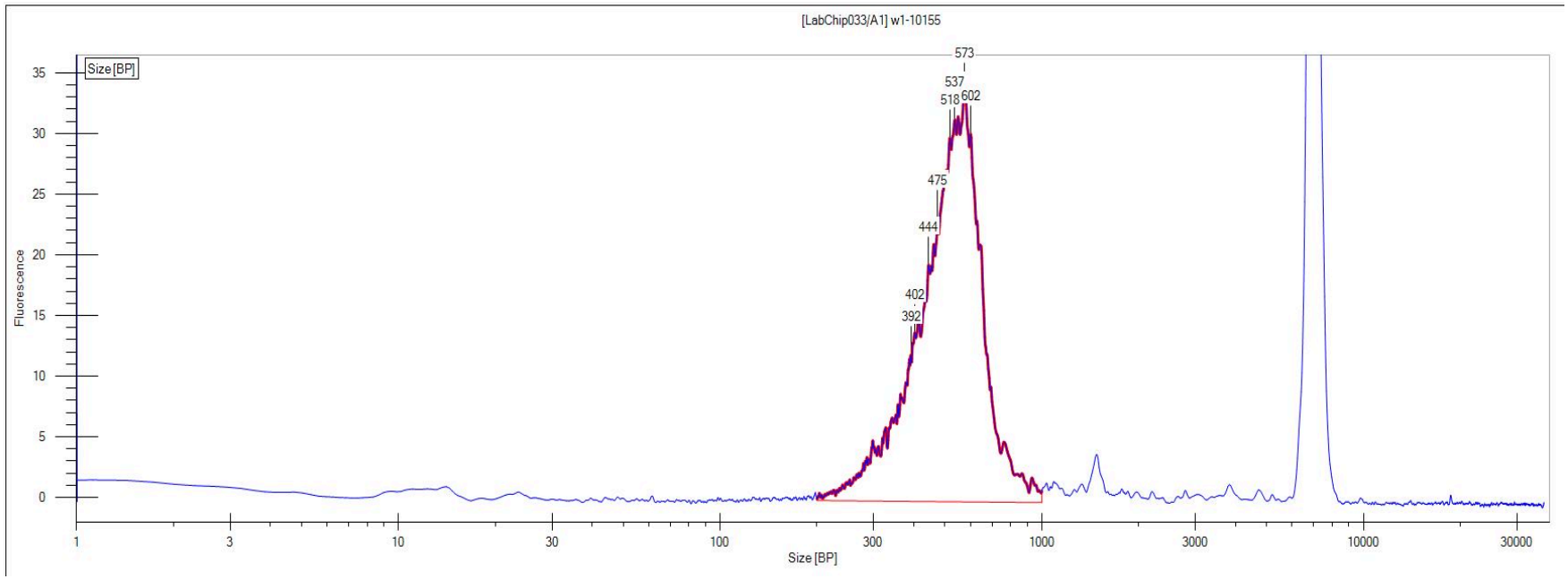


FIGURE 4-2. Fragment size by Agilent Technologies BioAnalyzer. The size distribution of DNA fragments plus adaptors, as measured by fluorescence.

4.3.6 Whole genome sequencing with Illumina Miseq

DNA libraries were then normalized to 10 nM and double-stranded DNA fragments were then denatured and hybridized (i.e., attached) to the flow cell. Using solid-phase bridge amplification, each fragment was then copied up to 1,000 times (134), forming millions of clonal clusters. With each cycle, a single nucleotide (A, T, C or G), complementary to the fragment being sequenced, was added by DNA polymerase. These nucleotides had fluorescently-labelled reversible terminators (134), which simultaneously prevent more than one nucleotide from being incorporated at once and facilitate identification. As each nucleotide was added, lasers were passed over the flow cell. This activated the fluorescent label, which was then detected and recorded (134). The amplification and clustering described previously is required to produce a signal of sufficient magnitude for detection. Each base was sequenced in succession; after the signal was recorded for a single nucleotide, the reversible terminator on this nucleotide was cleaved, allowing the next to be incorporated (134).

Sequencing was run in batched fashion. A single Illumina MiSeq run is generally performed on multiples of 24 samples, depending on the desired depth of coverage (defined as the average number of times each locus in the genome is sequenced). Though it is feasible to run the sequencer with fewer samples, the cost is equivalent whether 1 sample or 96 are run on the same plate (excluding expenses associated with DNA extraction and library preparation). Our 2x250 bp analysis required an average depth of coverage of at least 20x, which is thought to be sufficient to discern true variation from sequencing error (135). With a genome of 4.4 million bp for *M. tuberculosis*, this meant a maximum of 83 samples theoretically could have been run simultaneously (http://support.illumina.com/downloads/sequencing_coverage_calculator.html).

4.3.7 Sequence data storage and data sharing

Once sequencing was complete, reads from each isolate were saved in the form of 'fastq' files. These were available for download from Nanuq, a cloud server operated by the McGill University and Génome Québec Innovation Centre. All sequence data have been stored locally, and as a requirement for publication of these studies, was uploaded to the National Center for Biotechnology's Information Sequence Read Archive under Accession number SRP039605 (BioProject PRJNA240330).

4.3.8 Bioinformatics pipelines – from fastq to VCF

The bioinformatics pipelines employed in Manuscripts I and II have been described in detail in the corresponding supplementary methods (**Appendices 2-2** and **3-2**, respectively). Dr. Nicolas Radomski, a Post-Doctoral Fellow in the laboratory of Dr. Behr, was responsible for the steps prior to SNP calling for these 2 manuscripts.

After completing my first Objective, I dedicated the time to teach myself how to conduct bioinformatics for WGS data analysis. I subsequently developed a bioinformatics pipeline based on lessons learned from published manuscripts, reviewing "Best Practices" from the Broad Institute (for human genetic data), attending relevant conferences and discussing with experts at these forums. This pipeline has been utilized in Objective 4 and follows the basic workflow described in (136) (for human whole genome sequencing projects). Many of these steps were additionally recommended in a recent review of SNP calling methods for bacterial WGS data (137), further supporting this approach.

4.3.8.1 Trimming

Each base is given a quality score by the sequencing platform, reflecting the probability of an error at that locus. These scores, along with many others provided by bioinformatics tools, are Phred-scaled, where Phred = $-10*\log P_{error}$. Base qualities for each position in the sequenced reads were examined using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were then trimmed to remove residual adaptor sequences as well as low-quality bases (136) from the 3' ends (bases with Phred <30, corresponding to >1/1,000 probability of error).

4.3.8.2 Alignment

Reads were then aligned to the reference genome using the Burrows Wheeler Aligner (BWA) MEM algorithm, as this was shown to be more accurate than 6 other alignment tools and optimal for reads >70 bp in length (138). Once aligned, reads are stored in Binary Alignment/Map (BAM) format; an example of such an alignment, from a single isolate, is shown below.

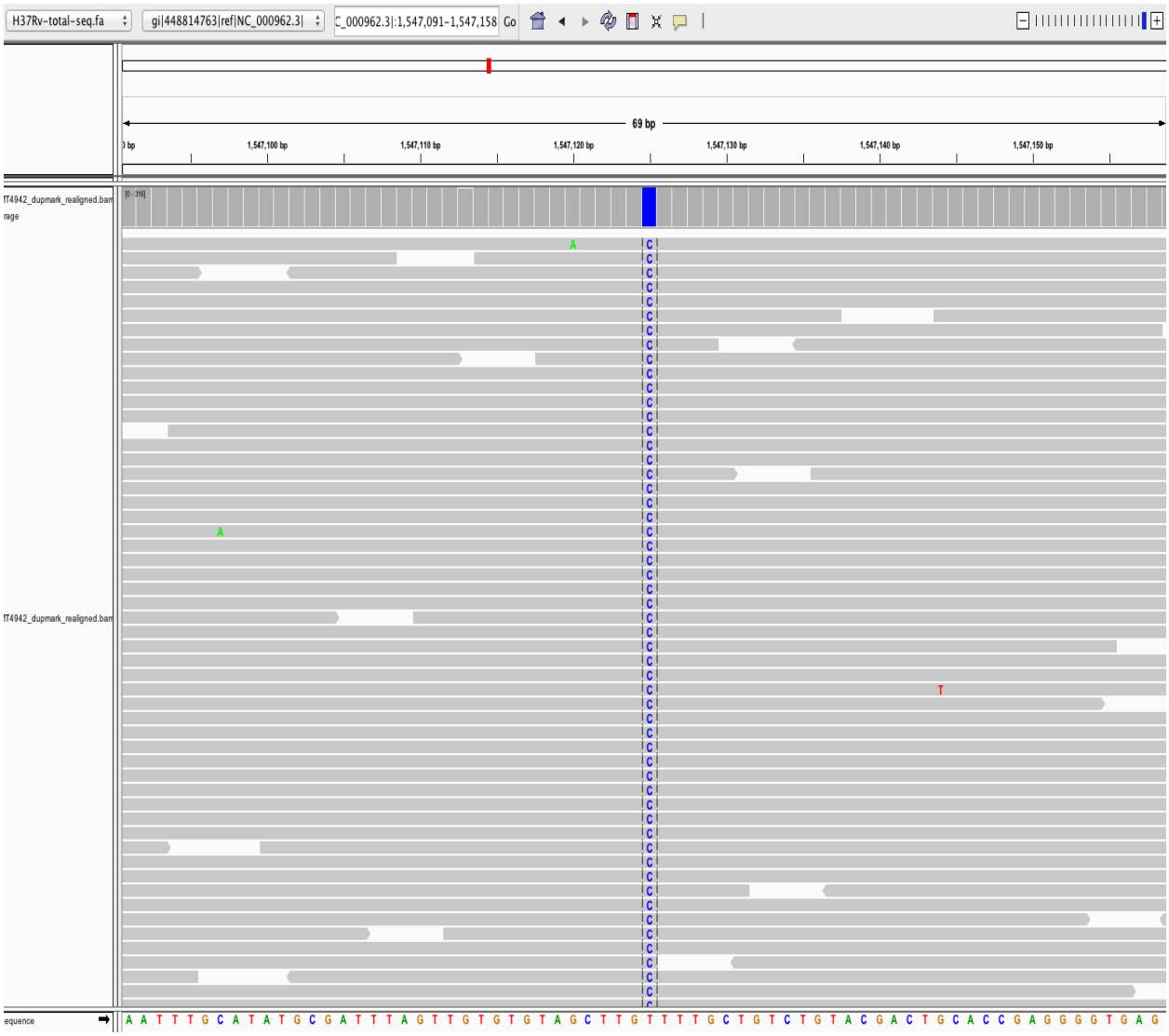


FIGURE 4-3. Screen shot of Integrative Genomics Viewer (139)**.** The alleles at each position in the H37Rv reference genome are indicated along the bottom. Grey bars represent individual reads, which have been aligned to the corresponding position in H37Rv. At any position, there can be 'reference' alleles (i.e., the same alleles as in H37Rv) and 'alternative' alleles (different alleles compared to H37Rv). Reference alleles are not indicated, while alternative alleles are represented by coloured letters corresponding to genotype. The alternative allele in blue is found in >50 reads at that locus, while in other loci, alternative alleles are found in only one read. While the former suggests a true SNP, the latter are more indicative of sequencing error.

Each read is assigned a mapping quality score based on the quality of its alignment. Like base quality scores, these are Phred-scaled, with lower scores indicating decreased confidence in the alignment (137). I excluded reads with mapping quality scores <30 using SAMtools (140). These

included reads that mapped ambiguously to >1 locus, which typically involve repetitive regions of the genome. PCR duplicates from the library preparation stage and optical duplicates, wherein the fluorescent signal from a single base is accidentally counted more than once, were also marked for exclusion ((136), done using Picard, available from http://broadinstitute.github.io/picard/); these can artificially amplify confidence in SNPs (137) or propagate errors in sequencing. Local realignment of reads around insertions and deletions was performed using GATK (141) to reduce potential alignment error related to these structural variants. This has been shown to greatly improve the positive predictive value of SNPs identified in downstream analyses (142).

Following these quality control steps, the percentage of total reads successfully aligned, the genome coverage (defined as the percent of the reference genome that has at least 1 read aligned to it) as well as the depth of coverage (i.e., the average number of reads aligned to any locus in the genome) were reviewed, and compared with other samples. Low values in any of these parameters would be a possible indication of contamination with another species, as the reads that fail to align are likely highly divergent from the reference.

4.3.8.3 SNP calling

To date, no studies have examined the sensitivity and specificity of SNP calling algorithms using bacterial genomes. Two studies (142, 143) using simulated or human genetic data were identified with depth of coverage similar to that obtained in bacterial studies. Overall, these studies found that sensitivity and specificity of the Genome Analysis ToolKit (GATK)'s Unified Genotyper ranged from 95.87-99.78% and 99.68-99.99%, respectively (142, 143). Comparing Unified Genotyper to other popular SNP callers (SAMtools, glf), Liu *et al.* (143) found that sensitivity was consistently higher for Unified Genotyper across a range of sequencing depth (4x, 10x, 20x). Using exome array as the gold standard, Unified Genotyper with multi-sample calling had an overall positive predictive value (PPV) of 99.3%, while SAMtools and GlfMultiples had PPVs of 97.8% and 97.9%, respectively (143). Similarly, Pirooznia *et al.* also found a higher PPV with Unified Genotyper compared to SAMtools, albeit PPV was lower compared to the other study, at 92.6% and 80.4%, respectively (142).

Therefore, to identify ('call') SNPs with respect to the reference genome, Unified Genotyper algorithm was used. Unified Genotyper utilizes a Bayesian genotype likelihood model, which is expressed as (141):

$$p(G|D) = \frac{p(G)p(D|G)}{p(D)},$$

where p represents a probability or probability density (if the variable is continuous);
D represents the data, i.e., the bases across each read at a given locus in the genome;
G represents the genotype;
p(G|D) is the posterior probability of the genotype;
p(G) is the prior probability of the genotype, based on previous knowledge about its prevalence in the population;
p(D|G) is the likelihood;
and p(D) is a normalizing constant, i.e., this does not vary across genotypes.

Diploid models have been used with (haploid) bacterial genomes to facilitate detection of mixed infection (e.g., (144, 145)); this would resemble a 'heterozygous' base call (146). Under a diploid model, there are 10 possible genotypes (A/A, T/T, C/C, G/G, A/T, A/C, A/G, T/C, T/G, C/G).

For each of these genotypes, the probability of the base identified in each read is calculated. This is repeated across all bases (i.e., across all reads in the 'pileup') for a given locus in the genome. The products of these probabilities equal the likelihood of the genotype. Mathematically,

$$p(D|G) = \prod_{b \in \text{pileup}} p(b|G),$$

where b is the base in that locus (141).

The probability of each base given a particular genotype is

$p(b|G) = ½\ p(b \mid A_1) + ½\ p(b \mid A_2)$ for each allele in $\{A_1, A_2\}$

where

$$p(b \mid A) = \begin{cases} \varepsilon/3 : b \neq A \\ 1-\varepsilon : b = A \end{cases}$$

and ε is the reversed Phred-scaled quality score for the base (141).

p(G) is specified by the software, and is influenced by whether the genotype is homozygous for either reference or alternative alleles, or heterozygous . Heterozygous genotypes have a prior probability of 0.001 or 1 per 1,000 bp (141).

In summary, given the prior probability of the genotype in the population, and the bases on each read at a specific locus, a posterior probability is determined for each possible genotype. Log-odds scores then are calculated by comparing the genotype with the greatest posterior probability to the second highest. If the log-odds score exceeds a set threshold, the genotype is reported, along with summary data, in Variant Call Format (VCF) (141).

4.3.8.4 Filtering SNPs

As base quality and mapping quality scores might not detect all sequencing or alignment errors, filtering is performed to ensure only high-confidence SNPs are included in analysis (135). SNPs were filtered using VCFtools (147) and SnpSift (148) based on Phred quality score, as well as minimum depth of coverage (137). SNPs were also assessed for strand bias, wherein all bases supporting the SNP were exclusively on forward or reverse reads; this is indicative of systematic sequencing error, rather than true variation (149). SNPs were also excluded if they were clustered within 12 bp of one another or within 20 bp of small insertions/deletions, as these are also more likely due to sequencing error. Furthermore, because of difficulty in accurately aligning to repetitive regions, SNPs found in the PE_PGRS and PPE genes of *M. tuberculosis* as well as transposable elements were excluded, as recommended by (100, 101).

It is important to note that, with the application of such filters to reduce false positive calls, true SNPs may be inadvertently overlooked (137). All phylogenetic trees based on these final SNP datasets were therefore validated by comparison to a deletion-based phylogeny constructed in Objective 2 (see the published supplemental data in **Appendix 3-2**).

4.3.8.5 Annotation

SNPs were annotated with their genomic locations (e.g., in gene X, intergenic, upstream or downstream of gene X, 3' or 5' untranslated region) using SnpEff (148). This software then predicts the functional impact of each SNP, based on these locations and the type of SNP. Nonsynonymous SNPs change the amino acid causing a change in the protein produced, and can therefore have high functional consequence, especially if they result in new start or stop codons. In contrast, synonymous SNPs (for the most part) are silent mutations. They are substitutions that do not change the amino acid, however, they may affect the expression level of the protein through regulatory effects.

4.3.9 Sanger Sequencing of SNPs

In addition to the filters applied above to reduce the risk of false positives, cluster-defining SNPs in *Rv0828c*, *carB*, *Rv1835c*, and *Rv3263* were also confirmed for 6 randomly-selected isolates from each cluster (IIIA, IIIB, IIIC) using Sanger sequencing (150). Primers were designed by Dr. Radomski to target each cluster-defining SNP. A minimum of 100 bp was required between the target SNP and both forward and reverse primers to ensure the target base was sequenced. Sequencing was performed at the McGill University and Génome Québec Innovation Centre.

4.3.10 Assessing for contamination and/or mixed infection

Contamination or mixed infection with species other than *M. tuberculosis* was evaluated using BLASTn (151) for a random sample of ~10,000 reads from each isolate.

This approach cannot discriminate mixed infection by different strains of *M. tuberculosis,* therefore a SNP-based approach was utilized. The distribution of heterozygous alleles was first reviewed across all isolates to assess for outliers, wherein mixed infection might be likely (123) (see published supplemental data, **Appendix 2-2**).

In addition to reviewing the distribution of heterozygous base calls, all loci defining clusters or subgroups in the 'outbreak' village were manually inspected to determine underlying base frequencies. Low-frequency bases (i.e., minority alleles that were not sufficient to classify the

base ‘heterozygous’) are typically indicative of sequencing/mapping error, but may also represent true variation, due to a subpopulation of bacteria within the host.

To investigate these two possibilities, all loci with >1 allele detected were reviewed manually in IGV and underlying base quality in each read was assessed. Where clinical epidemiologic data supported the possibility of mixed infection or micro-evolution and assessment of individual base calls did not rule this out, deep sequencing (re-sequencing to >150x depth of coverage) was performed. This approach has been recommended as an optimal way to detect mixed infection (152). The same SNP loci were then re-examined using these data.

4.4 Statistical modeling

The following section elaborates on the statistical modeling utilized in each Objective. While each manuscript clearly states the approaches used, the theoretical underpinnings of these have not previously been discussed.

4.4.1 Models of nucleotide substitution for inferring phylogenies

Throughout this thesis, maximum likelihood methods were used to estimate phylogenetic trees with the highest probability of yielding the included sequences (153). For this approach, it is necessary to first specify an underlying probability model of nucleotide substitution (154). Nucleotide substitution models describe the rate of change in bases at each locus over time, where each site is considered independent (155). A number of evolutionary models based on continuous time-reversible Markov processes have been developed to reflect different levels of complexity (154).

The choice of evolutionary model is critical, as incorrect models can lead to errors in tree topology (e.g., (156, 157)). In the accompanying analyses, 24 different model comparisons were performed to facilitate model selection, considering a combination of 6 different models of nucleotide substitution, uniform or gamma-distributed rates across sites, and invariant positions. The final models were selected based on the lowest Bayesian Information Criterion, using Molecular Evolutionary Genetics Analysis (158). These included the General Time Reversible (GTR, (159)) and the Tamura 3-parameter model (160). Considering the following matrix (155):

$$Q = \{qij\} = \begin{bmatrix} . & r_2\pi_C & r_4\pi_G & r_6\pi_T \\ r_1\pi_A & . & r_8\pi_G & r_{10}\pi_T \\ r_3\pi_A & r_7\pi_C & . & r_{12}\pi_T \\ r_5\pi_A & r_9\pi_C & r_{11}\pi_G & . \end{bmatrix}$$

In the above, each row (i) and column (j) is arranged in order of A, C, G and T; 'r_i' represents the rate of change from these alleles to the nucleotide indicated by the corresponding subscript. The GTR model assumes unequal base frequencies ($\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$) and estimates 6 different rate parameters for nucleotide substitution ($r_1\pi_A = r_2\pi_C$; $r_3\pi_A = r_4\pi_G$; $r_5\pi_A = r_6\pi_T$; $r_7\pi_C = r_8\pi_G$; $r_9\pi_C = r_{10}\pi_T$; $r_{11}\pi_G = r_{12}\pi_T$). The Tamura 3-parameter model is a less complex, nested model (160) which estimates both the transition-transversion rate ratio (i.e., the ratio of the substitution rates for purine-to-purine versus pyrimidine-to-pyrimidine) and the frequency of π_C, where $\pi_C = \pi_G$ and $\pi_A = \pi_T = (1 - \pi_C)/2$.

4.4.2 Nonparametric bootstrap

When Frequentist approaches are used for phylogenetics (e.g., maximum likelihood methods), nonparametric bootstrap is the most frequently employed method to assess confidence in tree topology. Nonparametric bootstrap performed by randomly sampling nucleotides (with replacement) from the original DNA sequences of each isolate. These random samples are then used to generate artificial, pseudo-sequences. Phylogenetic trees are produced, applying the same statistical procedure as previous, and concordance with the original tree is assessed (161). This process is repeated typically 100 to 1,000 times, with higher numbers of samples yielding higher precision. The distribution of these replicate trees around the observed data is considered an valid approximation of the sampling distribution of the observed data, i.e., it approximates the distribution that would have been obtained had sampling been repeated from the same underlying population (161). Each internal node in the original maximum likelihood tree is assigned a P value, which corresponds to the proportion of bootstrap replicate trees that supported the same phylogeny. Nodes with high bootstrap are considered to have higher confidence. While initially proposed as a measure of repeatability (102), the bootstrap P value is commonly used to assess accuracy of phylogenetic trees. Using simulations as well as real viral

DNA sequences, Hillis and Bull (162) illustrated that bootstrap P values >70% represent true phylogenetic relationships over >95% of the time, while P values >80% reflected nearly 100% accuracy.

4.4.3 Bayesian inference for molecular dating

Bayesian molecular dating was used in Objective 2 to address the public health concern that a new strain had been introduced and was transmitting throughout the North. In this study, isolates from 163 patients over 23 years were examined and two predominant sublineages were identified (the 'Major' and 'Minor' sublineage, responsible for 153 and 10 cases, respectively). Estimating the timing of introduction of these strains in the North required extrapolating back from the sequences of these isolates to their most recent common ancestor (MRCA, representing the imputed ancestral sequence) and estimating the time at which divergence from this ancestor occurred (tMRCA). Bayesian molecular dating represents the predominant approach for estimating divergence times using next-generation sequencing data (163).

4.4.3.1 Bayesian molecular dating and Markov chain Monte Carlo methods

Bayesian statistics utilize a continuous form of Bayes' Theorem. In phylogenetics, this is expressed as:

$$f(\theta|D) = \frac{f(\theta) * p(D|\theta)}{p(D)}$$

where D is the sequence data from the included isolates;
θ represents the unknown parameters in the model (including the parameters in the model of nucleotide substitution, the molecular clock and parameters of the demographic model);
f(θ|D) is the posterior distribution;
f(θ) is the prior distribution, which is a joint prior over all parameters in θ;
p(D|θ) is the likelihood;
P(D) is a constant, representing the marginal likelihood of D (modified from (164)).

Ignoring the constant p(D), as this does not have to be calculated,

$f(\theta|D) \propto f(\theta) * p(D|\theta)$

The principles behind Bayesian inference are as follows (informed by Bayesian statistics courses taught by Dr. Lawrence Joseph in the Department of Epidemiology at McGill University): in

brief, prior distributions represent what is already known about the parameters of interest, including previous evidence from the literature about their distributions (e.g., their means and standard deviations). By specifying an independent prior for each parameter in θ, one can explicitly incorporate this evidence into the current analysis. The strength of a prior is described by the variance assigned to its prior density, with smaller variance suggesting greater support for the proposed means. If there is weak or no evidence from the literature to inform the prior, a non-informative prior can be used, such that the probability is equally distributed over a wide range of values (165). An example of a non-informative prior is the following:
$X \sim \text{uniform}[-\text{inf},+\text{inf}]$
where X represents a random variable and equal probability is given to any value between negative infinity and positive infinity.

Once individual priors are specified, they are multiplied together to form a joint prior distribution. These will then be 'updated' based on the data, as expressed by the likelihood function $p(D|\theta)$. The likelihood function is often the same as function utilized in Frequentist statistics. In phylogenetics, this is frequently Felsenstein's likelihood (153, 164).

To derive the posterior density, the joint prior distribution and the likelihood are multiplied. If non-informative priors are used, the posterior distribution will be based solely on the data (expressed in the likelihood function). The resulting 95% credible intervals obtained will be approximately numerically equivalent to Frequentist 95% confidence intervals, provided the sample size is large (165). However, if an informative prior is utilized, the weight given to the data in calculating the posterior is dependant both on the strength of the prior and the sample size of the data. As sample size increases, the weight given to the likelihood (i.e., the data) compared to the prior increases (165).

In the case of Bayesian molecular dating, the age of the MRCA is inferred. However, as in many other Bayesian analyses, solving the exact posterior distribution is analytically intractable (161). Therefore, Markov chain Monte Carlo methods (MCMC) are utilized to obtain random samples from the posterior distribution. As explained in (164), a starting position ('state') is chosen for all parameters in θ. A new state is then proposed, given the previous state, through the proposal

distribution $q(\theta'|\theta)$. If the new state explains the observed data with higher probability than the previous state, it is accepted. If the new state has lower probability than the previous state, a random number is drawn from a uniform distribution between (0, 1). If this random number is less than or equal to $f(\theta'|D)/f(\theta|D)$, the new state is accepted. If this number is greater than $f(\theta'|D)/f(\theta|D)$, the new state is rejected. Therefore, if the new state is significantly worse, it will be rejected (164).

In conditions where the proposal distribution is not symmetric (i.e., $q(\theta'|\theta) \neq q(\theta|\theta')$), reversibility is maintained via the Hastings ratio; if a proposed state in this instance is worse, the random number is now compared with $(f(\theta'|D)/f(\theta|D))*(q(\theta|\theta')/q(\theta'|\theta)$ (164).

As each state depends on the previous, with only small changes between them, adjacent states are highly correlated. In order to account for this, the chain length of the MCMC (i.e., the number of states generated) must be large enough to ensure independent samples can be taken. In Bayesian Evolutionary Analysis by Sampling Trees (BEAST, (166)), an effective sample size (ESS) of 100 independent samples from the posterior distribution for each parameter is considered adequate, with higher ESS indicating higher confidence.

Samples should only be taken after the Markov chain has reached stationarity, i.e., when it has converged on the target posterior distribution. Samples taken prior to reaching convergence should be discarded. This is called the 'burn-in' period (167). As the initial values used to start the MCMC may be highly divergent from the true posterior distribution or relatively close, the number of states needed to reach convergence varies. A burn-in of 10% is often considered adequate, but this should be assessed using trace plots of the values taken for each continuous parameter across all samples (**Figure 4-4**) (164, 168). Those with similar mean and variance throughout indicate stationarity. Adequate mixing (i.e., how well the MCMC moves through the sample space) can also be assessed, to ensure the chain is sampling a range of values from the posterior distribution.

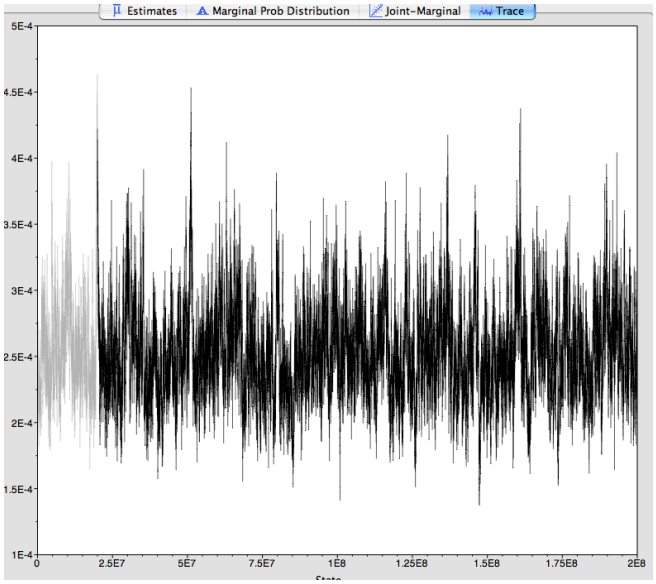


FIGURE 4-4. Values for a single parameter, across all samples in Tracer, (169)**.** This is an example of adequate mixing and stationarity of the MCMC, with ESS >200. Burn-in is sufficient (in grey), as there are no overall upward or downward trends visible.

Once convergence and adequate mixing are confirmed, mean/median values and 95% highest posterior density (HPD) intervals can be obtained for the posterior probability of each parameter. The latter represents the most narrow interval containing X% of the posterior probability (164).

4.4.3.2 Choosing prior distributions

Prior distributions on the parameters represented by θ are informed by previous knowledge, including results from similar studies. When such knowledge is limited, non-informative prior distributions are used.

Among parameters requiring specification of a prior distribution is the molecular clock, i.e. the rate of substitution over time. As mentioned in **Chapter 3**, previous studies using different datasets have estimated a molecular clock of ~0.5 SNPs per genome per year or $1.3x10^{-7}$ per site (53, 101). However, as indicated in Bryant *et al.* (106), there can be substantial variation around this estimate. Therefore, to assess the validity of applying a strict (i.e., a single) clock across our whole dataset, a likelihood ratio test was performed in MEGA (158). The probability of a strict clock was rejected at $p < 0.05$. Based on this, a uncorrelated relaxed log-normal clock (170) prior

was used, allowing for variation across branches of the phylogenetic tree. A rate was then estimated, overall and for both Major and Minor sublineages. The initial value was set at $1.3x10^{-7}$ per site to reduce time to convergence of the MCMC, as - even if variable - clock rates would be expected to be in this numerical range.

In order to estimate divergence times, a tree prior must also be specified. Tree priors are population demographic models, used to describe changes in size of a bacterial population over time. In all analyses, a coalescent tree prior was used, which assumes that all isolates in the sample will ultimately 'coalesce', i.e., they all share a common ancestor at some point in history. This is ideal when samples arise from the same population (164) and by using such a prior, this accounts for the dependence between the genetically-related sampled isolates (168). Model selection was performed using the posterior-simulation based analogue of Akaike's Information Criterion (AICM, (171)). Among parametric models of population demographics, a coaelescent prior with a constant population size was selected based on the lowest AICM (168). This model was used for the first two MRCA analyses, with a non-informative Jeffery's prior for the population size parameter. A third analysis applied a non-parametric coalescent model (the Bayesian skyline) (172) to infer the estimated population size as a function of time, without specification of a functional form for population size.

Default, non-informative priors were used for the rates estimated by the model of nucleotide substitution (GTR, as previous).

4.4.3.3 Constructing time trees

Along with all parameters, phylogenetic trees and branch lengths were also sampled from the posterior distribution. These phylogenetic trees are directed, meaning all isolates are thought to descend from a single imputed ancestor (the MRCA). This is the 'root' of the tree, with relative time of 1.0. Time 0 is the most recently-collected isolate. Extrapolating from the genetic diversity in the contemporaneous sequences, ancestral nodes are identified where at least 1 isolate diverged from another. There are n-1 such ancestral nodes, where n represents the number of isolates (168). Divergence times are calculated relative to the root MRCA for each node. To

convert divergence times to absolute time (i.e., into years rather than relative to the root MRCA), additional dating information is required.

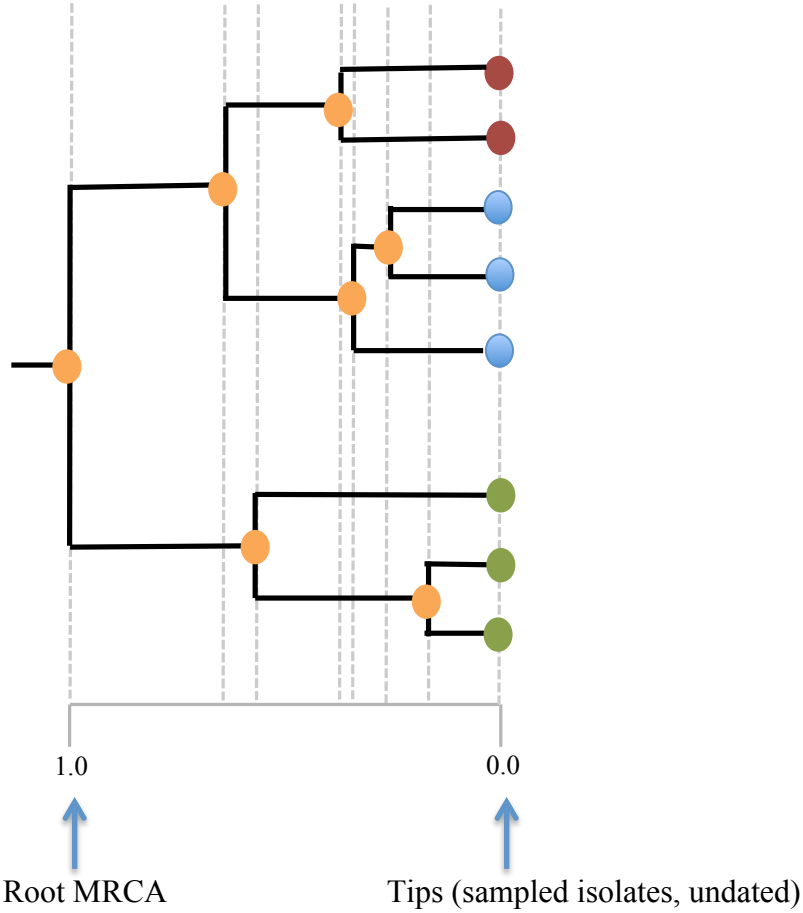


FIGURE 4-5. Example of a relative time tree. The root of the tree is assigned a divergence time of 1.0, with all isolates at the tips assigned a relative time of 0. Different clusters of isolates are indicated in red, blue and green. The most recent common ancestors between all isolates are inferred. These nodes are indicated in orange and assigned a relative time compared to the root.

In our analyses, relative time was converted into absolute time by calibrating the ancestral node of all isolates in the Major sublineage. This calibration was based on the previously reported substitution rate of 0.5 SNPs per genome per year and the extremities of reported 95% confidence intervals (0.3-0.7 SNPs per genome per year) (101, 105). To assess the robustness of our findings and influence of this prior, two additional analyses did not include a calibration node, relying solely on the dates isolates were sampled (i.e., the years they were collected from the patients) to convert relative into absolute divergence times.

Based on the sampled trees, a maximum clade credibility tree was generated. This is the tree that has the highest product of all posterior clade probabilities (173). Median divergence times (i.e., tMRCAs) and HPD intervals were reported for all nodes with posterior density >0.8 (174), which corresponds to nodes that are found in at least 80% of sampled trees. This ensured high confidence in the distal branches of phylogeny and therefore the corresponding tMRCAs.

CHAPTER 5. OBJECTIVE 1 – Manuscript I

Lee RS, Radomski N, Proulx J-F, Manry J, McIntosh F, Desjardins F, Soualhine H, Domenech P, Reed MB, Menzies D, Behr MA. Reemergence and re-amplification of tuberculosis in the Canadian Arctic. *J Infect Dis* 2015;211(12):1905-1914

5.1 Preamble

In a single year, one village in Nunavik had 50 cases of microbiologically-confirmed TB (population 933). With an incidence of TB higher than anywhere else in the world for that year, and an extraordinary attack rate of ~20%, this outbreak was alarming in many ways to both the community and local health care providers. A massive public health response helped identify, diagnose and treat patients as rapidly as possible, as well as start patients on LTBI prophylaxis.

Several of the authors herein, myself included, were directly involved in the public health response to this 'outbreak'. Following this, once no further waves of cases were being diagnosed, we collaborated with the village council to investigate this extraordinary event.

The first aim of this thesis work, to resolve transmission networks in this village, is described in the following manuscript. The published reprint of this manuscript is enclosed in **Appendix 2-1**. This is followed by the accompanying supplementary data, which includes detailed methods, in **Appendix 2-2**.

5.2 Manuscript I

Re-emergence and Amplification of Tuberculosis in the Canadian Arctic

Robyn S. Lee, BSc[1–3†], Nicolas Radomski, PhD[3,†], Jean-Francois Proulx, MD[4],Jeremy Manry, PhD[2,3,5], Fiona McIntosh, BSc[3], Francine Desjardins, DTL[6], Hafid Soualhine, PhD[7], Pilar Domenech, PhD[3], Michael B. Reed, PhD[2,3], Dick Menzies, MD[3,8],Marcel A. Behr, MD[2,3*]

1. McGill University, Department of Epidemiology and Biostatistics
2. McGill International TB Centre
3. The Research Institute of the McGill University Health Centre
4. Nunavik Regional Board of Health and Social Services
5. McGill University, Departments of Medicine and Human Genetics
6. McGill University Health Centre, Royal Victoria Hospital
7. Laboratoire de Santé Publique du Québec
8. Montreal Chest Institute, Respiratory Epidemiology and Clinical Research Unit

† These authors contributed equally to this manuscript.

Full addresses of authors:

- Robyn S. Lee, Montreal General Hospital, 1650 Cedar Avenue, RS1-105, Montreal, PQ, Canada, H3G 1A4
- Nicolas Radomski, Montreal General Hospital, 1650 Cedar Avenue, RS1-105, Montreal, PQ, Canada, H3G 1A4
- Jean-Francois Proulx, Régie régionale de la santé et des services sociaux Nunavik , C.P. 900, Kuujjuaq, PQ, Canada, J0M 1C0
- Jeremy Manry, Montreal General Hospital, 1650 Cedar Avenue, L1-413, Montreal, PQ, Canada, H3G 1A4
- Fiona McIntosh, Montreal General Hospital, 1650 Cedar Avenue, RS1-111, Montreal, PQ, Canada, H3G 1A4
- Francine Desjardins, McGill University Health Centre, Royal Victoria Hospital, L5·09, Montreal, Canada, PQ, H3A 1A1
- Hafid Soualhine, Laboratoire de Santé Publique du Québec, 20045 Sainte-Marie Ch, Sainte-Anne-de-Bellevue, PQ, Canada, H9X 3R5
- Pilar Domenech, Montreal General Hospital, 1650 Cedar Avenue, RS1-133, Montreal, PQ, Canada, H3G 1A4
- Michael B. Reed, Montreal General Hospital, 1650 Cedar Avenue, RS1-133, Montreal, PQ, Canada, H3G 1A4
- Dick Menzies, Montreal Chest Institute, Room K 1·24, Montreal, PQ, Canada, H2X 2P4

*Correspondence and requests for reprints to:

Dr. Marcel A. Behr

Phone: 514 934-1934 (42815)

Fax: 514 934-8423

Email: marcel.behr@mcgill.ca

Running head: Amplification of Tuberculosis in the Arctic

[1] No authors have any commercial or other associations that may pose a conflict of interest.

[2] This study was funded by the Canadian Institutes of Health Research (MOP-125858).

Abstract

Between November 2011 and November 2012, a Canadian village of 933 persons had 50 culture-positive cases of tuberculosis, with 49 sharing the same genotype.

Methods

We performed Illumina-based whole-genome sequencing on Mycobacterium tuberculosis isolates from this village, during and before the outbreak. Phylogenetic trees were generated using the maximum likelihood method.

Results

Three distinct genotypes were identified. Strain I (n = 7) was isolated in 1991-1996. Strain II (n = 8) was isolated in 1996-2004. Strain III (n = 62) first appeared in 2007 and did not arise from strain I or II. Within strain III, there were 3 related but distinct clusters: IIIA, IIIB, and IIIC. Between 2007 and 2010, cluster IIIA predominated (11 of 22 vs. 2 of 40; P <.001), whereas in 2011-2012 clusters IIIB (n = 18) and IIIC (n = 20) predominated over cluster IIIA (n = 11). Combined evolutionary and epidemiologic analysis of strain III cases revealed that the outbreak in 2011-2012 was the result of ≥6 temporally staggered events, spanning from 1 reactivation case to a point-source outbreak of 20 cases.

Conclusions

After the disappearance of 2 strains of *M. tuberculosis* in this village, its reemergence in 2007 was followed by an epidemiologic amplification, affecting >5% of the population.

Introduction

Between November 2011 and November 2012, there were 50 cases of microbiologically proven active tuberculosis in an Arctic village in Nunavik, Québec. With a population of only 933, the incidence of culture-confirmed tuberculosis was >5% of the community for that year - 1000 times the overall Canadian incidence. This outbreak occurred in a setting with a very low prevalence of human immunodeficiency virus infection and no previous resistance to antituberculosis drugs, leading to concern in the populace of a newly emerged hyper-virulent strain of *Mycobacterium tuberculosis*.

As part of the response to the outbreak, the Nunavik Regional Board of Health and Social Services (NRBHSS) conducted extensive contact investigations of all newly diagnosed active tuberculosis cases, including household and social contacts. During this response, it was observed that many persons had contacts with multiple tuberculosis cases, indicating that it would be extremely difficult to identify transmission links using standard epidemiologic methods. An alternative approach would involve molecular typing of patient isolates.

In work published elsewhere, a combination of classic molecular epidemiology tools (restriction fragment length polymorphism [RFLP] and mycobacterial interspersed repetitive units [MIRUs]) revealed extremely limited bacterial diversity in this region, both within and across villages [1]. One potential interpretation of these findings is that this represents ongoing transmission. However, an alternative hypothesis is that patients share similar bacterial genotypes due to ancestry. With the advent of whole-genome sequencing (WGS), a higher-resolution molecular epidemiologic tool [2–6], it is now possible to test whether bacteria that are otherwise indistinguishable indicate recent transmission of *M. tuberculosis*. Furthermore, because WGS provides information on lineage-specific polymorphisms, this genotyping method can also determine whether a new, potentially more virulent *M. tuberculosis* strain had been introduced into this community.

To address these 2 questions, we conducted WGS on *M. tuberculosis* isolates from this village. To validate WGS data in this setting, we tested epidemiologically unrelated isolates from other villages of the same region, over 6 years. Then, to situate WGS data from 2011 to 2012 in the

context of a village with high rates of tuberculosis over many years, we extended our analysis to the 2 decades before the outbreak. In this setting with limited variability by conventional genotyping modalities, WGS provided improved analytic resolution, revealing the disappearance, reemergence, and amplification of *M. tuberculosis* over time.

Methods

Study Population

Nunavik, the arctic region of Québec, spans 443 685 km^2 and comprises 14 Inuit communities. The outbreak village, henceforth denoted village K, is >150 km from the nearest village, with no road connecting the communities.

Bacteria

Specimens from tuberculosis suspects in Nunavik are processed at the mycobacteriology laboratory of the McGill University Health Centre (MUHC). Culture-positive isolates are forwarded to the reference laboratory, Laboratoire de Santé Publique du Québec, for drug susceptibility testing. These laboratories provided isolates for the years 1991–2012.

Genomics

DNA extraction [7] and WGS have been described elsewhere [8], with details in the **Supplementary Data**. In brief, *M. tuberculosis* isolates were sequenced using the MiSeq 250 System (Illumina). Reads with a minimum length of 50 base pairs (bp) were retained and deposited in the National Center for Biotechnology Information's Sequence Archive (accession No. SRP039605, i.e. BioProject PRJNA240330). After alignment to the H37Rv reference genome (accession No. NC_000962.3), single-nucleotide polymorphisms (SNPs) were identified using a Bayesian likelihood model (Unified Genotyper; Genome Analysis ToolKit, version 2.7.4); SNPs with a minimum Phred score >50 were retained (where Phred is $-10 \cdot \log_{10} P_{error}$). Phylogenetic analysis was done using Molecular Evolutionary Genetics Analysis (MEGA, version 5, [9]), with the number of differences method used to compute evolutionary distance [10]. Maximum likelihood trees were generated using the model of nucleotide substitution that yielded the lowest Bayesian information criterion (Tamura 3-parameter model, [11]). As a sensitivity analysis, we also generated maximum likelihood trees using the Jukes–Cantor model

[12].

Validation of SNP Threshold for Recent Transmission

Given the limited genetic diversity in Nunavik [1], we evaluated the lowest SNP threshold that could occur in the absence of transmission. To do so, we sequenced *M. tuberculosis* isolates from cases residing in other villages of Nunavik (2006–2012). Contact investigation data were obtained from the NRBHSS. Case pairs without epidemiologic links who resided in different villages were designated as improbable transmission, and the SNPs between these case pairs were compared.

Application of WGS to Village K

The SNPs between village K isolates were identified, including those from cases diagnosed during the 20 years before the outbreak. Phylogenetic trees were generated while blinded to epidemiologic data.

Clinical Epidemiologic Analysis Combined With WGS

For the outbreak, clinical epidemiologic data were collected by clinical staff in village K. Links between cases were identified using a database of all household and named contacts. Using date of diagnosis/treatment initiation, symptoms, sputum smear status, and cavity on chest radiograph as indicators of contagion [13], we looked for potential index cases in each cluster. For the years preceding the outbreak, epidemiologic data for cases from 2007 to 2010 were provided by the NRBHSS. Smear microscopic results were obtained from the MUHC laboratory.

Statistical Analysis

A 2-sample z test and the exact binomial test were used to compare proportions. The F* test for samples with unequal variance was used to compare the number of pairwise SNPs within clusters. Analyses were conducted using Stata software (version 11, StataCorp 2009).

Ethical Approval

Ethical approval was obtained from the McGill University Faculty of Medicine's institutional review board and the NRBHSS. Individual patient consent was not required, but the study was

done in collaboration with the village K council.

Results

The Outbreak

Between November 2011 and November 2012, there were 50 microbiologically confirmed cases of tuberculosis in village K. There were no cases between January and October of 2011. All cases were pulmonary, with no instances of tuberculosis meningitis or disseminated disease. Seven of the 50 cases were diagnosed based on symptoms. Of the remaining 43 cases, 40 were found to have active disease during contact investigation, and 3 developed tuberculosis after a documented positive tuberculin skin test conversion; 1 had refused isoniazid and the other 2 demonstrated low adherence. The epidemiologic links between cases were highly complex (**Figure 5-1**). All cases except one shared the same MIRU pattern; RFLP provided similar resolution (**Supplementary Figure 1**).

Tuberculosis in Village K Over 22 Years

Between 1991 and 2012 (i.e., including the outbreak year), 82 cases of culture-positive tuberculosis were diagnosed in village K (**Figure 5-2**), yielding an average annual incidence of >450 per 100 000 (population denominators from Statistics Canada). The majority of cases were male (47 of 82), with a median age of 22 years (interquartile range, 16–35 years), consistent with the age distribution of this population [14].

Of the 82 confirmed cases in village K, 80 (97.6%) had isolates available for genotyping, 78 of which provided high-quality WGS data: 49 of 50 outbreak isolates, 14 of 15 isolates from 2007 to 2010, and all 15 isolates from 1991 to 2004 (there were no cases in 2005–2006). Average genome coverage among the 78 isolates was 99.7% (standard deviation [SD], 0.11%), with an average depth of coverage of 42× (SD, 13). The majority of Phred scores were between 500 and 1000 for SNPs, indicating minimal ascertainment bias, and there was no evidence supporting infection with multiple *M. tuberculosis* strains (**Supplementary Figure 2**). A review of isolates without SNPs between them revealed that the specimens were processed in separate batches, arguing against laboratory cross-contamination.

Validation of SNP Threshold for Recent Transmission

WGS was successful for 42 of 45 cases in other villages of Nunavik (2006–2012). Consistent with our observation of limited genetic diversity in this region, the 631 "improbable transmission" case pairs from other villages of Nunavik were separated by as few as 2 SNPs, but none were separated by 0 or 1 SNP (**Supplementary Figure 3**). From this finding, supported by studies published elsewhere, we defined a new cluster when a group of isolates shared ≥2 of the same SNPs compared with the reference group.

Application of WGS to Village K

The SNPs from all isolates of village K were used to infer maximum likelihood trees, with the bootstrap consensus tree from 1000 replicates shown in **Figure 5-3** [11, 15]. Results were robust to use of an alternate model of nucleotide substitution (unpublished data). All isolates were lineage 4 (Euro-American, with the reported 7-bp deletion in the *pks15/1* gene) [16], and 3 distinct strains were evident, designated strains I, II, and III (Figure 5-3). Neither strain I nor strain II gave rise to strain III; strain I has 16 unique SNPs not seen in strain III, whereas strain II has 18 unique SNPs plus a 1102-bp deletion (2 963 340–2 964 352) that is intact in strain III isolates.

Strain I predominated for 6 years (n = 7; 1991–1996), then disappeared. Strain II predominated for 9 years (n = 8; 1996–2004), then disappeared (**Figure 5-2**). Strains I and II were unique to village K. Strain III was first detected in village K in 2007, though it was subsequently found in 2 cases diagnosed in other villages. One of these cases was a child adopted from village K to another community, and the other was an adult who had been a close family contact of a smear-positive case in village K before developing active tuberculosis the following year.

Within strain III, 3 clusters were observed, designated IIIA, IIIB, and IIIC (**Figure 5-4**). Cluster IIIA isolates (n = 22) had the reference alleles for the genes *carB*, *Rv3263*, *Rv0828c*, and *Rv1835c*. Cluster IIIB isolates (n = 20) had cluster-defining SNPs in *carB* and *Rv3263* but were wild type for *Rv0828c* and *Rv1835c*; cluster IIIC isolates (n = 20) had cluster-defining SNPs in *Rv0828c* and *Rv1835c* but were wild-type for *carB* and *Rv3263*. Of the 3 clusters, IIIC had the least bacterial diversity (mean pairwise SNP difference between isolates, 1.7 [95% confidence interval, 1.5–1.8] within IIIA, 1.6 (1.4–1.8) within IIIB, and 0.4 (0.3–0.5) within IIIC; $P < .001$).

Clinical Epidemiologic Analysis Combined With WGS

Whereas WGS alone revealed 3 different clusters (IIIA, B, C), further analysis in conjunction with epidemiologic data identified more complex transmission networks over time, with ≥6 distinct subgroups from 2011 to 2012 (**Figure 5-5**, across the bottom). Cluster IIIA was first seen in 2007–2008 and was initially divided into 2 groups—those with the C allele in *mce1B* (n = 4) and those with an alternative T allele in this gene (n = 18).

Between 2011 and 2012, there were 11 cluster IIIA isolates. One had the C allele in *mce1B* and was from a familial contact of previous cases whose isolates had the same genotype in 2008, suggestive of an isolated reactivation event. The 10 remaining isolates had the T allele in *mce1B*. Two of these isolates also had an alternative C allele in *Rv0331*. In this latter subgroup, 1 case was diagnosed in November 2011 and had smear-positive (3+) cavitary disease (MT-5531), while the other was a household contact. The remaining 8 IIIA isolates were first observed in May 2012. Within this subgroup, there were 3 smear-positive cases (4+ for MT-3074, 3+ for MT-3341, and 2+ for MT-3673) diagnosed in June 2012 plus 5 more cases diagnosed at about the same time or soon afterward. Nearly all secondary cases were friends or family, with no obvious trend in locations of contact. Thus, the 11 cluster IIIA isolates from 2011 to 2012 are unlikely to represent a single transmission event, because ≥2 discrete transmission chains plus 1 isolated reactivation event are better supported by the combined genetic and epidemiologic data.

Cluster IIIB was first seen in 2009 and had the reference mce1B C allele, plus cluster-defining SNPs in *carB* and *Rv3263*. In 2011–2012, there were 18 cluster IIIB isolates. Five of these had an alternative C allele in *fadE4*, and the other 13 had the reference A allele at this position. The former subgroup was first seen in December 2011, when a single case was diagnosed with smear-positive (4+) cavitary disease (MT-504). The remaining 4 cases with this genotype were teenagers with shared attendance at the same "gathering house," a venue of socialization identified by public health during the outbreak. The latter subgroup (with the reference A allele in *fadE4*) was first detected 3 months later, in March 2012. Although it is possible that MT-504 had a mixed infection and contributed to both subgroups, we also note that cases with the alternative C allele were diagnosed months before those with the reference A allele. Moreover,

the group of 13 cases with the reference A allele included a patient with smear-positive (3+) cavitary disease diagnosed in May 2012 (MT-2474) who had definitive epidemiologic links to 4 of the remaining 12 cases. The combination of WGS and epidemiology together suggest that the 18 cluster IIIB isolates from 2011 to 2012 represent ≥2 transmission chains.

Cluster IIIC was not seen in the community before 2012. The first case was diagnosed in January 2012 with sputum smear–positive (3+) cavitary disease (MT-0080). Fifteen of the remaining 19 cases were epidemiologically linked to this case (4 household contacts, 3 friends, and 8 contacts at gathering houses). This putative source reported symptoms for 4 months before diagnosis, possibly explaining the large number of IIIC cases observed early in 2012 (8 additional cases in January–February 2012 and 3 in March–April). Of these cases, 2 were smear positive (2+ for MT-1838 and 2+ for MT-2151). Hence, some of the remaining cases with diagnoses between May and November 2012 may have been infected by these secondary cases. These data suggest that cluster IIIC represents, at a minimum, 1 discrete transmission chain.

The epidemiologic curve of the outbreak shows, at the village level, a bimodal distribution of cases diagnosed over time (**Figure 5-6A**). When outbreak cases were stratified by the aforementioned subgroups, the bimodal distribution was largely attributable to differences in the temporal presentation of the different clusters and their subgroups (**Figure 5-6B**). When examining the contact data on the most transmissible cases in each of the subgroups, we can tabulate the number of household and non-household contacts who developed active tuberculosis with the same genotype. As seen in **Table 5-1**, of named household contacts who developed tuberculosis, 56% shared the same genotype as the epidemiologically identified source. In contrast, among non-household contacts who developed tuberculosis, only 19% shared the same genotype as their putative source, which was no better than chance alone (exact binomial for comparison to 1/6, given 6 subgroups; P = .32).

Discussion

Using WGS, we have been able to reveal the complexity of tuberculosis control in a unique environment, where there is virtually no loss to follow-up and little to no in- or out-migration. On the scale of decades, 2 dominant strains have disappeared, not to be seen again after 1996 and

2004. Unfortunately, the reemergence of tuberculosis in or around 2007 was followed by a series of secondary and tertiary cases, culminating in an explosion of tuberculosis cases in 2011–2012. Whereas WGS alone revealed 3 clusters in the 2011–2012 outbreak, the combination of WGS with epidemiologic data allowed us to resolve this into a minimum of 6 events—5 transmission chains and 1 isolated case of reactivation. Together, these findings suggest that (1) even a single reactivation event can lead to numerous cases in this community and (2) the outbreak of 2011–2012 was not a single, rare occurrence but rather multiple smaller concurrent events. This suggests that this community is highly vulnerable to tuberculosis outbreaks, such that ongoing surveillance and vigilance against tuberculosis are warranted.

Our analysis of the outbreak leads us to several important conclusions. First, the outbreak was not due to the introduction of a new *M. tuberculosis* lineage. The isolates circulating in 2011–2012 differed by a maximum of 8 SNPs from those already present in 2007, and both IIIA and IIIB cases were documented in the years before the outbreak. Although we cannot exclude the possibility that the 2 nonsynonymous SNPs in strain IIIC affect bacterial fitness or virulence, this strain was responsible for less than half of the outbreak cases. It is therefore unlikely that these few mutations, on their own, accounted for the dramatic case rate of 2012. Rather, our findings suggest that the 2011–2012 outbreak involved the expansion of extant bacteria, consistent with a historical study of tuberculosis in Western Canada [17].

Second, both the WGS data and the clinical/epidemiologic data point to multiple transmission events, rather than a single outbreak. Although it remains possible that a single patient harbored a diversity of strains [18] and was therefore the sole source, such an explanation is neither likely nor necessary to explain the outbreak. Within a few years of the introduction of strain III, there were highly contagious carriers of each of IIIA, IIIB, and IIIC, each with epidemiologic links to multiple contacts sharing the same genotype. The knowledge that there are 3 clusters (IIIA, IIIB, and IIIC) in combination with epidemiologic data has also helped identify a case of exogenous reinfection that would otherwise have been overlooked given the absence of MIRU variability. In addition, the cluster-defining SNPs of IIIA/B/C are now being used to investigate the sources of 2013–2014 cases and to distinguish relapse from reinfection in recurrent cases.

Finally, whereas MIRU and RFLP of this community would suggest that there is, and has been, ongoing transmission in this village for decades [1], WGS data challenge this interpretation. Strains I and II disappeared in 1996 and 2004, respectively, before the introduction of strain III. Given that strain III was first seen in village K and differs from strains I and II by approximately 40 SNPs, the most plausible explanation is a single reactivation case due to an organism acquired in the same village, decades before the period sampled. The majority of adults in the village have positive tuberculin skin test results, and many have chronic pulmonary diseases, so it is possible that one such individual developed transmissible disease without medical suspicion of tuberculosis, leading to the introduction of strain IIIA in 2007.

It remains unclear why this population was at such a high risk after the reappearance of tuberculosis in 2007. Given that one of the potential source cases in the outbreak presented to the clinic 4 months after symptom onset, patient delay may be a considerable factor in this population. Furthermore, although the majority of household contacts with tuberculosis shared the same genotype as the most transmissible cases within each subgroup, 44% of these household contacts did not, supporting the findings of Verver *et al.* [19] that in an environment with high tuberculosis transmission, the traditional stone-in-pond principle may not suffice for identifying and interrupting transmission. As implemented in 1954 in Alaska [20], community-wide interventions, such as chest radiographic screening, may be needed to halt tuberculosis transmission in this setting. BCG vaccination was already reinstituted in the village in response to this outbreak after its cessation in 2005.

The primary limitation of this study is the relatively small sample size of the subgroups revealed by WGS. Despite the extraordinary incidence of disease, there was insufficient power to conduct a rigorous statistical comparison between cases in the different transmission chains. Another potential limitation is that we were unable to sequence 4 of 82 isolates. However, because we successfully sequenced 95% of all isolates from village K between 1991 and 2012, there is minimal risk of sampling bias. Finally, from a public health perspective, we were unable to identify a single, unifying cause of the 2011–2012 outbreak; this is not surprising, however, given that in-depth analysis revealed the outbreak was in fact due to ≥6 epidemiologically distinct events.

There are a number of important strengths of this study. The unique environment, with nearly all isolates sharing the same MIRU pattern, provided the opportunity to examine how limited classic genotyping methods can actually be. We have demonstrated that although isolates in a transmission chain share the same MIRU, the converse does not necessarily hold true—a fact that may have important implications for public health investigation of MIRU-defined clusters. The analysis by WGS of a single geographically isolated village provided an unexpected opportunity to witness both the disappearance and reemergence of tuberculosis over time. Isolates sequenced had a minimum coverage of 21x, and 97% of the SNPs identified had a Phred score of >100, equivalent to a 1 in 10^{10} chance of error. The results for the outbreak obtained using the maximum likelihood method were concordant with both the previously established rate of mutation of *M. tuberculosis* [3, 5, 21, 22] and independent results from Nunavik outside village K (**Supplementary Figure 3**). Our phylogeny also proved robust to use of an alternate model of nucleotide substitution. We obtained independent confirmation of the 4 cluster-defining SNPs for clusters IIIA, IIIB, and IIIC using Sanger sequencing, and a previous study by Domenech et al [8] also showed very low false-positive rates using the same WGS pipeline. Finally, detailed clinical epidemiologic data were available for all cases, facilitating the verification of transmission identified by WGS.

In summary, the use of WGS permitted a fine-level analysis of an ongoing tuberculosis epidemic in this vulnerable population. The reappearance of *M. tuberculosis* was followed several years later by an epidemiologic amplification, leading to a multipronged outbreak affecting >5% of the population. Further consideration of the potential mechanisms of tuberculosis spread in this village, and other communities in Nunavik, is warranted to derive strategies to help these and other vulnerable communities control and ultimately eliminate tuberculosis.

Acknowledgments

We thank the village council and the residents of Kangiqsualujjuaq for their collaboration and engagement in this study. We also thank the staff of Centre Local de Services Communautaires Palaqsivik for their dedicated care of patients and contacts during the outbreak. Thanks to Genevieve de Bellefeuille, BSc (Agence de la Santé et des Services Sociaux de l'Estrie), for her

hard work collecting clinical and epidemiologic data during the outbreak; Isabelle Rocher, MSc (Institut National de Santé Publique du Québec), for assisting with data entry while working clinically during the outbreak; and Erwin Schurr, PhD (The Research Institute of MUHC), for his input on the genetic analysis. We also thank the NRBHSS for detailed collection of clinical and epidemiologic data for the duration of the study.

Financial support

This work was supported by the Canadian Institutes of Health Research (MOP No. 125858).

Potential conflicts of interest.

All authors: No reported conflicts.

FIGURES

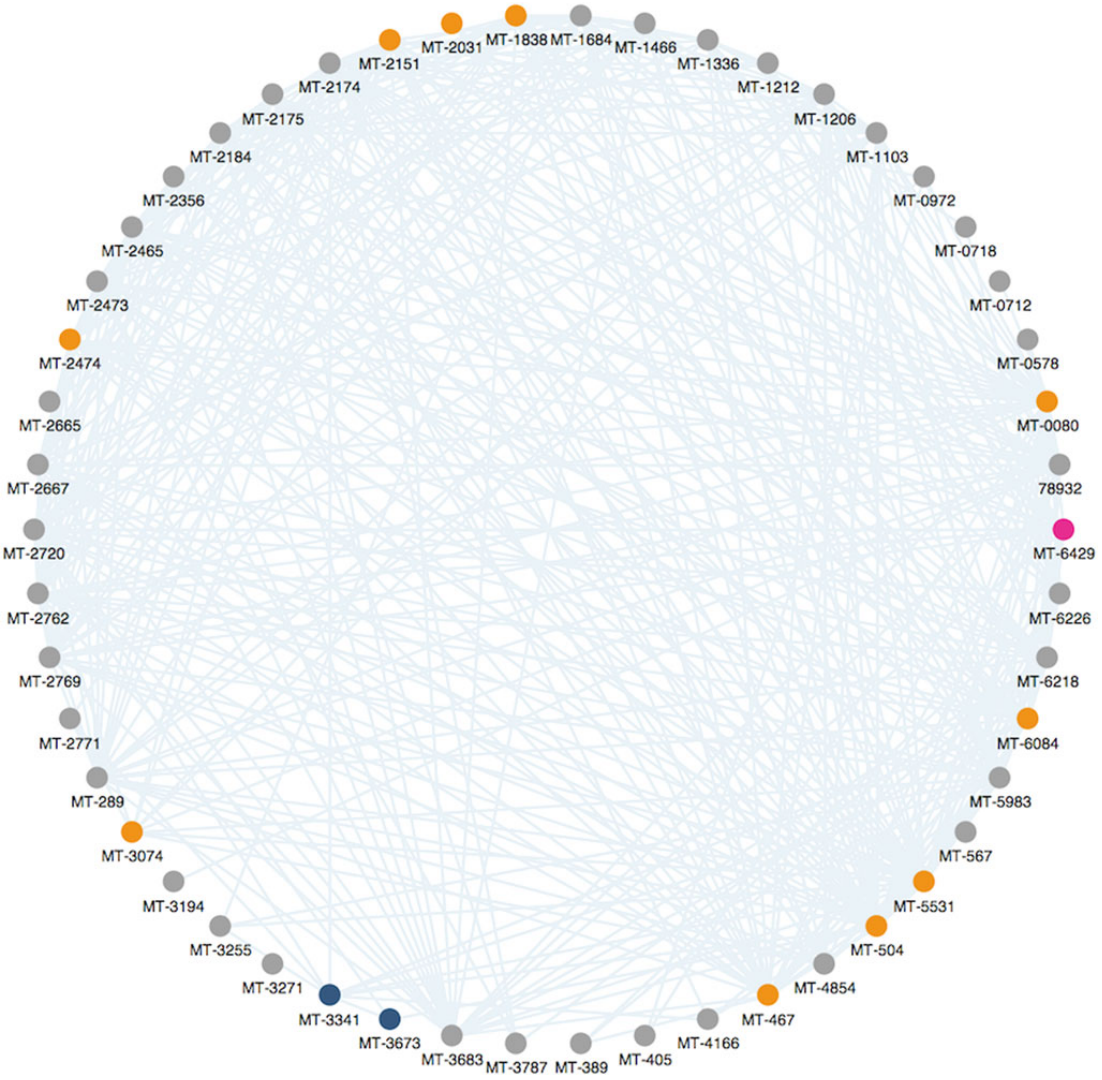


FIGURE 5-1. Epidemiologic links between outbreak cases.
Links between household/named contacts, as well as shared attendance (or residence) of community "gathering houses" identified by contact investigation are indicated. Orange circles represent sputum smear-positive cavitary cases; navy circles, sputum smear-positive non-cavitary cases; pink circles, sputum smear-negative cavitary cases; gray circles, sputum smear-negative non-cavitary cases.

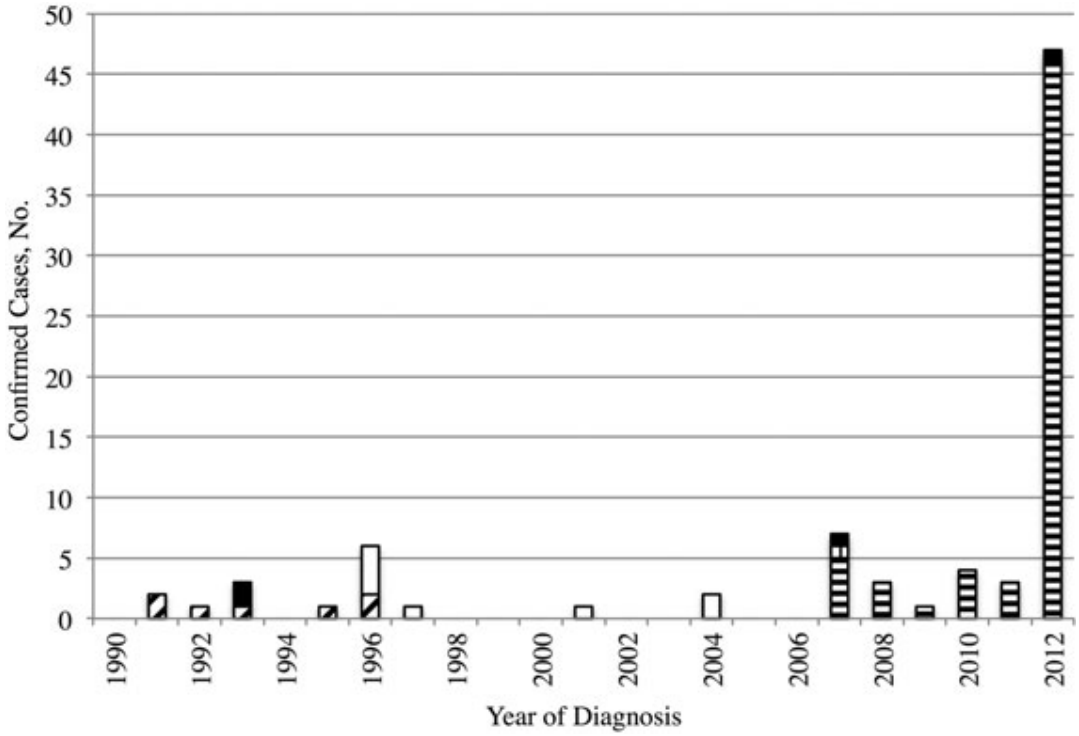


FIGURE 5-2. Microbiologically confirmed tuberculosis in village K (1990– 2012).
The numbers of confirmed tuberculosis cases reported in village K from 1990 to 2012 are shown by year of diagnosis. Strains of isolates are indicated, as identified by whole-genome sequencing (WGS): diagonal stripes indicate strain I; solid white, strain II; horizontal stripes, strain III; and vertical stripes, not clustered; solid black represent isolates for which WGS was not available.

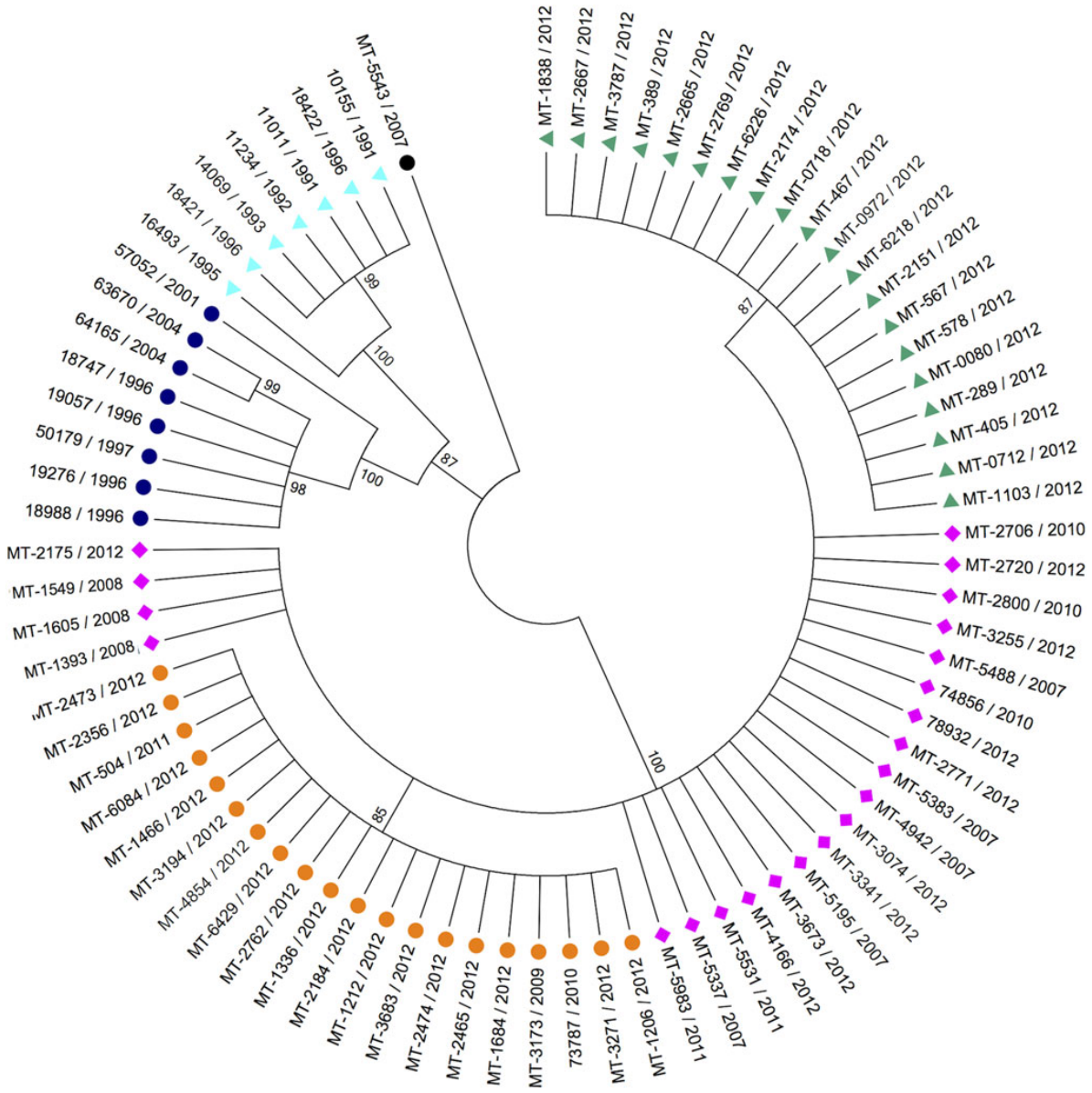


FIGURE 5-3. Bootstrap consensus tree of Mycobacterium tuberculosis isolates from village K.

The evolutionary history was inferred by using the maximum likelihood method based on the Tamura 3-parameter model [11] and a bootstrap consensus tree was generated with 1000 replicates [15]. The percentage of trees in which the associated genome clustered together is shown next to the branches. Branches reproduced in <80% of bootstrap replicates are collapsed. Initial trees for the heuristic search were obtained by applying the neighbor-joining method to a matrix of pairwise distances estimated using the maximum composite likelihood approach. The analysis involved 78 genomes compared with the H37Rv reference genome. Light blue triangles represent strain I isolates; dark blue circles, strain II; pink diamonds, strain III, cluster A; orange circles, strain III, cluster B; green triangles, strain III, cluster C; black circle, not clustered.

						Strain				
						I	II	III		
								Cluster		
Position in H37Rv	Gene code	Gene name	Reference allele	Alternative allele	Mutation			IIIA	IIIB	IIIC
1331500	Rv1188	*Rv1188*	A	G	Nonsynonymous	G	G	A	A	A
1567583	Rv1392	*metK*	C	T	Synonymous	T	T	C	C	C
2022484	Rv1784	*Rv1784*	C	T	Synonymous	T	T	C	C	C
22398	Rv0018c	*ppp*	T	G	Nonsynonymous	G	T	T	T	T
42401	N/A	N/A	G	C	Intergenic	C	G	G	G	G
383629	Rv0315	*Rv0315*	C	T	Nonsynonymous	T	C	C	C	C
428744	Rv0355c	*PPE8*	A	G	Nonsynonymous	G	A	A	A	A
558829	N/A	N/A	C	G	Intergenic	G	C	C	C	C
1011972	Rv0908	*ctpE*	G	A	Nonsynonymous	A	G	G	G	G
2115139	Rv1866	*Rv1866*	G	A	Nonsynonymous	A	G	G	G	G
3251386	Rv2932	*ppsB*	C	T	Synonymous	T	C	C	C	C
3280352	Rv2940c	*mas*	C	A	Nonsynonymous	A	C	C	C	C
3367489	Rv3009c	*gatB*	T	C	Synonymous	C	T	T	T	T
3368965	Rv3010c	*pfkA*	T	G	Nonsynonymous	G	T	T	T	T
3910540	Rv3492c	*Rv3492c*	G	A	Synonymous	A	G	G	G	G
4093885	Rv3652	*PE_PGRS60*	C	T	Nonsynonymous	T	C	C	C	C
30642	Rv0026	*Rv0026*	C	T	Synonymous	C	T	C	C	C
403587	Rv0338c	*Rv0338c*	G	A	Nonsynonymous	G	A	G	G	G
677113	N/A	N/A	C	T	Intergenic	C	T	C	C	C
775381	Rv0675	*echA5*	A	C	Nonsynonymous	A	C	A	A	A
852682	Rv0758	*phoR*	T	G	Nonsynonymous	T	G	T	T	T
1310662	Rv1179c	*Rv1179c*	C	T	Nonsynonymous	C	T	C	C	C
2472939	Rv2208	*cobS*	A	C	Synonymous	A	C	A	A	A
2706379	Rv2408	*PE24*	C	T	Synonymous	C	T	C	C	C
3137638	Rv2831	*echA16*	C	G	Nonsynonymous	C	G	C	C	C
3206398	Rv2896c	*Rv2896c*	A	C	Nonsynonymous	A	C	A	A	A
3503781	Rv3137	*Rv3137*	G	A	Nonsynonymous	G	A	G	G	G
3973994	Rv3535c	*Rv3535c*	C	T	Synonymous	C	T	C	C	C
3975488	Rv3537	*Rv3537*	C	T	Synonymous	C	T	C	C	C
3997479	Rv3557c	*Rv3557c*	G	A	Synonymous	G	A	G	G	G
4218303	Rv3773c	*Rv3773c*	G	A	Nonsynonymous	G	A	G	G	G
567913	N/A	N/A	T	C	Intergenic	T	T	C	C	C
600950	Rv0509	*hemA*	G	A	Synonymous	G	G	A	A	A
809416	Rv0712	*Rv0712*	A	G	Nonsynonymous	A	A	G	G	G
1360496	Rv1217c	*Rv1217c*	C	A	Nonsynonymous	C	C	A	A	A
1837767	Rv1633	*uvrB*	G	A	Synonymous	G	G	A	A	A
1917190	Rv1692	*Rv1692*	G	A	Nonsynonymous	G	G	A	A	A
1973649	Rv1747	*Rv1747*	C	T	Nonsynonymous	C	C	T	T	T
2228900	N/A	N/A	T	G	Intergenic	T	T	G	G	G
2501354	Rv2227	*Rv2227*	G	A	Nonsynonymous	G	G	A	A	A
2623818	N/A	N/A	G	A	Intergenic	G	G	A	A	A
2760387	Rv2458	*mmuM*	A	C	Synonymous	A	A	C	C	C
2783792	Rv2477c	*Rv2477c*	G	T	Nonsynonymous	G	G	T	T	T
2976814	Rv2653c	*Rv2653c*	G	A	Synonymous	G	G	A	A	A
3108453	Rv2800	*Rv2800*	T	C	Nonsynonymous	T	T	C	C	C
3374067	Rv3014c	*ligA*	A	C	Nonsynonymous	A	A	C	C	C
3430937	Rv3066	*Rv3066*	G	A	Synonymous	G	G	A	A	A
3467000	Rv3097c	*lipY*	T	A	Nonsynonymous	T	T	A	A	A
3599965	Rv3224	*Rv3224*	G	A	Nonsynonymous	G	G	A	A	A
3772940	Rv3360	*Rv3360*	C	T	Nonsynonymous	C	C	T	T	T
4232293	Rv3785	*Rv3785*	C	G	Nonsynonymous	C	C	G	G	G
4287890	Rv3822	*Rv3822*	G	A	Synonymous	G	G	A	A	A
4409769	Rv3921c	*Rv3921c*	G	C	Nonsynonymous	G	G	C	C	C
1558108	Rv1384	*carB*	C	T	Synonymous	C	C	C	T	C
3644579	Rv3263	*Rv3263*	G	A	Nonsynonymous	G	G	G	A	G
921390	Rv0828c	*Rv0828c*	T	C	Nonsynonymous	T	T	T	T	C
2082436	Rv1835c	*Rv1835c*	C	T	Nonsynonymous	C	C	C	C	T

FIGURE 5-4. Strain and cluster-defining single-nucleotide polymorphisms (SNPs) for strains I, II, and III.
Strain and cluster-defining SNPs shown. Reference and alternative alleles are highlighted in white and gray, respectively. From a progenitor strain, strains I and II have evolved distinctly from strain III, itself further subdivided into clusters IIIA, IIIB, and IIIC. Alleles in the genes *Rv0828c*, *carB*, *Rv1835c*, and *Rv3263* (H37Rv loci 1 558 108, 3 644 579, 921 390 and 2 082 436, respectively) were confirmed by Sanger sequencing for 6 isolates from each of clusters IIIA, IIIB, and IIIC.

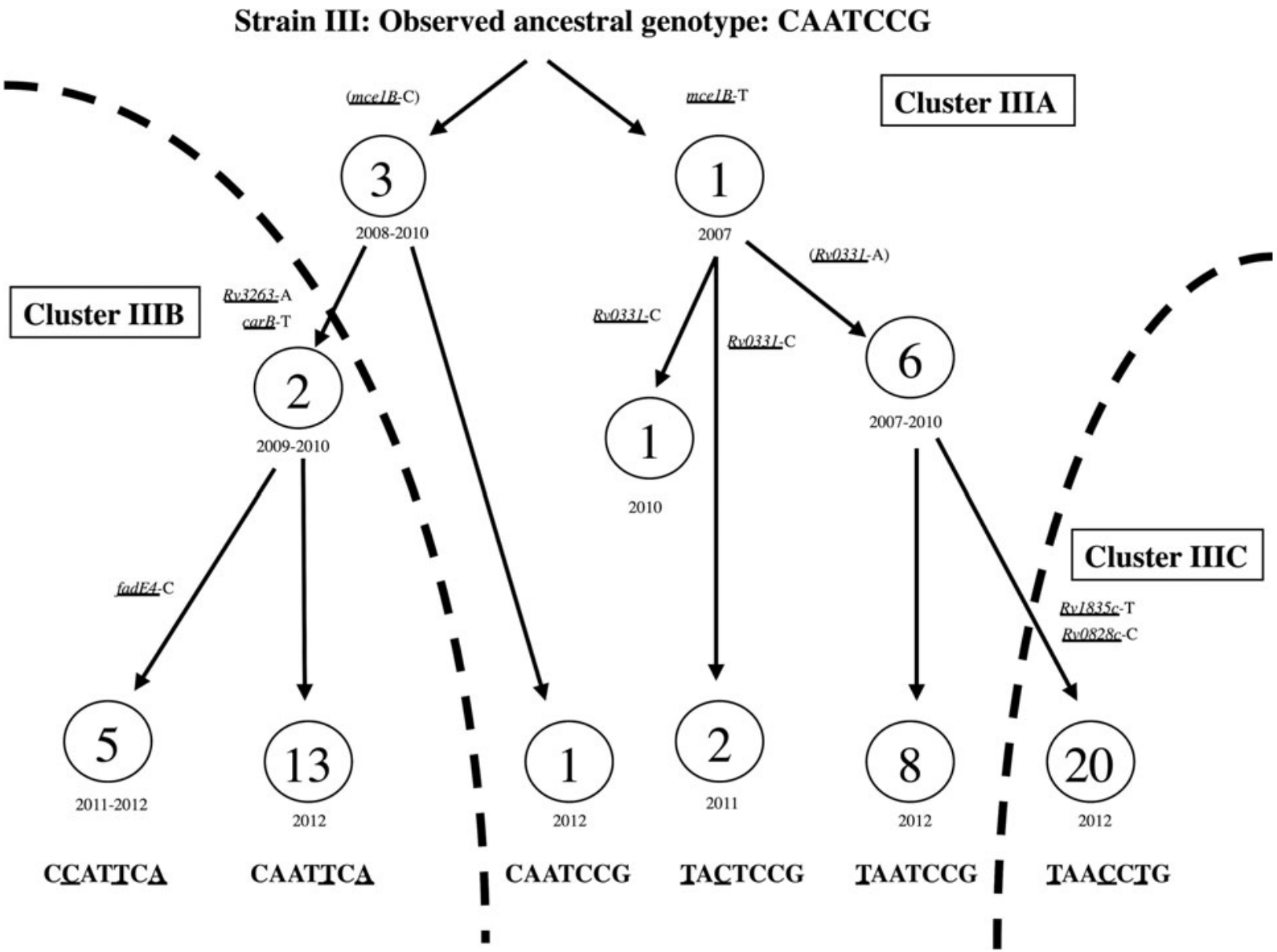

Strain III: Observed ancestral genotype: CAATCCG
(mce1B-C)
mce1B-T
Cluster IIIA
3
2008-2010
1
2007
(Rv0331-A)
Cluster IIIB
Rv3263-A
carB-T
Rv0331-C
Rv0331-C
2
2009-2010
1
2010
6
2007-2010
Cluster IIIC
fadE4-C
Rv1835c-T
Rv0828c-C
5
2011-2012
13
2012
1
2012
2
2011
8
2012
20
2012
CCATTCA
CAATTCA
CAATCCG
TACTCCG
TAATCCG
TAACCTG

FIGURE 5-5. The microevolution of strain III in village K over time, involving a total of 7 single-nucleotide polymorphisms (SNPs).
Numbers in circles indicate numbers of cases at each stage of evolution. The years of all isolates in each group are indicated below the circles, with time scaled from the top (2007) to the bottom (2012). Arrows indicate bacterial microevolution, and SNPs are identified by the gene name and the corresponding allele; to highlight certain lineages with the reference allele, the gene name and allele are in parentheses. Starting with the ancestral genome (top), cluster IIIA had 2 initial subgroups, one with the reference allele C at *mce1B* (4 cases; middle panel, left) and the other with alternative allele T at *mce1B* (18 cases; middle panel, right). Within the latter 18 cases, there were 2 subgroups: 3 with an additional variant at *Rv0331* and 15 that retained the reference allele, A. Cluster IIIB (bottom left) was derived from strain IIIA with the *mce1B* C reference allele, with 2 additional mutations (in *Rv3263*, *carB*). A subgroup of 5 had an additional variant in *fadE4*. Cluster IIIC (bottom right) was derived from cluster IIIA, with the alternative T allele at *mce1B* and 2 additional mutations (in *Rv1835c*, *Rv0828c*). At the bottom, the concatenated genotype for the 7 SNPs is presented for each of the 6 subgroups identified during the outbreak year.

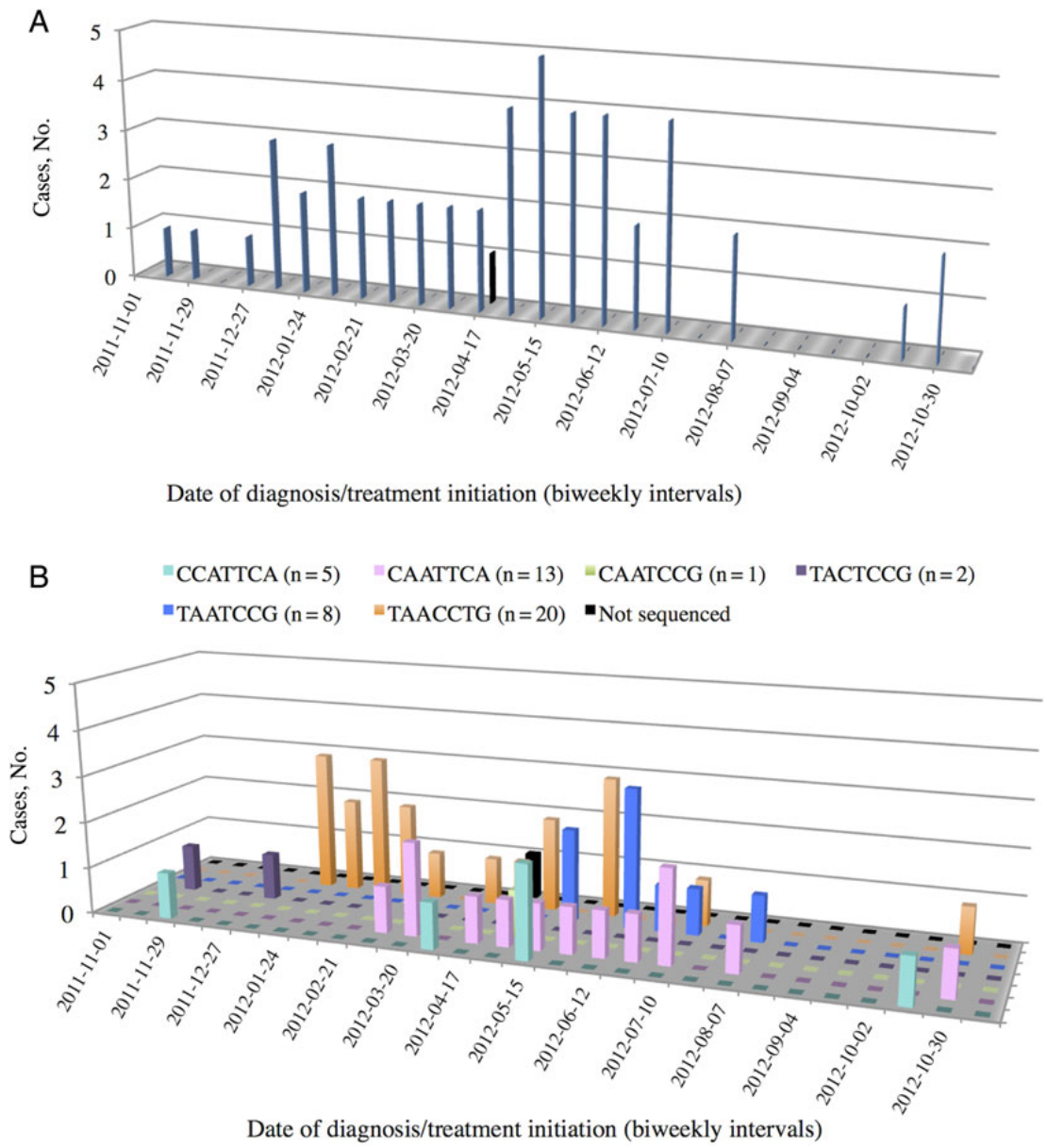


FIGURE 5-6. Epidemiologic curves of the outbreak.
A, Overall. The numbers of cases during the outbreak are shown by date of diagnosis (year-month-date). Blue represents isolates for which whole-genome sequencing (WGS) was successful; black, isolates without WGS. There were no cases before November in 2011. B, Epidemiologic curve of the outbreak, stratified by WGS/epidemiologic subgroup. The numbers of cases during the outbreak are shown by date of diagnosis (year-month-date), in biweekly intervals. Cases are stratified by subgroup genotype, as indicated.

TABLES

TABLE 5-1. Household and Social Contacts With Active Tuberculosis of the Same Genotype for Each Smear-Positive Case by WGS Epidemiologic Subgroup [a]

	Date of diagnosis			Contacts with same genotype/total contacts, No. (%)[b]	
Subgroup by WGS and Epidemiology	1st case	Smear positive cases	Smear grade	Household contacts	Social contacts
IIIA, n=1	May 2012	-	-	0/0 (-)	0/18 (0)
IIIA, n=2	Nov. 2011	Nov. 2011	3+	1/4 (25)	0/30 (0)
IIIA, n=8	May 2012	Jun. 2012	4+	0/0 (-)	3/10 (30)
		Jun. 2012	3+	1/1 (100)	4/9 (44)
		Jun. 2012	2+	1/1 (100)	1/3 (33)
IIIB, n=5	Dec. 2011	Dec. 2011	4+	0/0 (-)	4/32 (13)
		Oct. 2012	3+	0/3 (0)	2/22 (9)
IIIB, n=13	Mar. 2012	May 2012	2+	2/2 (100)	3/21 (14)
IIIC, n=20	Jan. 2012	Jan. 2012	3+	3/3 (100)	12/31 (39)
		Apr. 2012	2+	1/3 (33)	5/20 (25)
		May 2012	2+	1/1 (100)	8/23 (35)
Total (n/N)				10/18 (56)	42/219 (19)[c]

Abbreviation: WGS, whole-genome sequencing. [a] Smear positive was defined as 1+ or higher, except the first subgroup comprised only 1 person, who had smear-negative disease. [b] Because different sources named the same contacts, the denominators of contacts who developed active tuberculosis exceed the number of unique cases in the year. [c] A 2-sample z test was used to assess difference in proportions ($P < .001$).

References

1. Nguyen D, Proulx JF, Westley J, Thibert L, Dery S, Behr MA. Tuberculosis in the Inuit community of Quebec, Canada. Am J Resp Crit Care Med 2003; 168:1353–7.

2. Gardy JL, Johnston JC, Ho Sui SJ, *et al*. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N Engl J Med 2011; 364:730–9.

3. Walker TM, Ip CL, Harrell RH, *et al*. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. Lancet Infect Dis 2013; 13:137–46.

4. Schurch AC, Kremer K, Daviena O, *et al*. High-resolution typing by integration of genome sequencing data in a large tuberculosis cluster. J Clin Microbiol 2010; 48:3403–6.

5. Roetzer A, Diel R, Kohl TA, *et al*. Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiological study. PLoS Med 2013; 10:e1001387.

6. Kato-Maeda M, Ho C, Passarelli B, et al. Use of whole genome sequencing to determine the microevolution of *Mycobacterium tuberculosis* during an outbreak. PLoS One 2013; 8:e58235.

7. van Soolingen D, Hermans PW, de Haas PE, Soll DR, van Embden JD. 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. J Clin Microbiol 1991; 29:2578–86.

8. Domenech P, Rog A, Moolji JUD, et al. The origins of a 350-kilobase genomic duplication in *Mycobacterium tuberculosis* and its impact on virulence. Infect Immun 2014; 82:2902–12.

9. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 2011; 28:2731–9.

10. Nei M, Kumar S. Molecular evolution and phylogenetics. New York, NY: Oxford University Press, 2000.

11. Tamura K. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. Mol Biol Evol 1992; 9:678–87.

12. Jukes TH, Cantor CR. Evolution of protein molecules. New York, NY: Academic Press, 1969:21–132.

13. Centers for Disease Control. Guidelines for the investigation of contacts of persons with infectious tuberculosis. MMWR Recommend Rep 2005; 54:1–62.

14. Statistics Canada. 2012. Kangiqsualujjuaq, Quebec (Code 2499090) and Quebec (Code 24) (table). Census Profile. 2011 Census. Statistics Canada Catalogue no. 98-316-XWE, Ottawa. Released October 24, 2012. http://www12.statcan.gc.ca/census-recensement/2011/dp-pd/prof/index.cfm?Lang=E. Accessed 10 December 2014.
15. Felsenstein J. Confidence-limits on phylogenies: an approach using the bootstrap. Evolution 1985; 39:783–91.
16. Comas I, Coscolla M, Luo T, *et al*. Out-of-Africa migration and neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. Nat Genet 2013; 45:1176–82.
17. Pepperell CS, Granka JM, Alexander DC, *et al*. Dispersal of *Mycobacterium tuberculosis* via the Canadian fur trade. Proc Natl Acad Sci USA 2011; 108:6526–31.
18. Perez-Lago L, Comas I, Navarro Y, *et al*. Whole genome sequencing analysis of intrapatient microevolution in *Mycobacterium tuberculosis*: potential impact on the inference of tuberculosis transmission. J Infect Dis 2013; 209:98–108.
19. Verver S, Warren RM, Munch Z, *et al*. Proportion of tuberculosis transmission that takes place in households in a high-incidence area. Lancet 2004; 363:212–4.
20. Grzybowski S, Styblo K, Dorken E. Tuberculosis in Eskimos. Tubercle 1976; 57(suppl 4):S1–58.
21. Ford CB, Shah RR, Maeda MK, *et al*. *Mycobacterium tuberculosis* mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. Nat Genet 2013; 45:784–90.
22. Bryant JM, Schürch AC, van Deutekom H, *et al*. Inferring patient to patient transmission of *Mycobacterium tuberculosis* from whole genome sequencing data. BMC Infect Dis 2013; 13:110.

5.3 Additional unpublished analyses

To assess for mixed infection or within-host microevolution which could affect inferences of transmission, the SNPs that distinguished the 6 different subgroups were investigated manually for all isolates in strain III from 2007-2012.

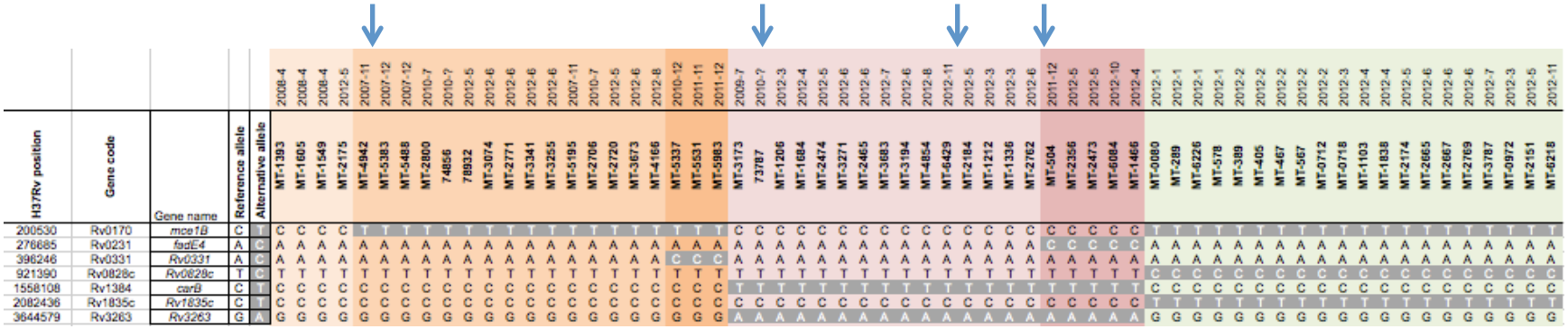


FIGURE 5-7 Cluster and subgroup – defining SNP loci in strain III. Alleles are shown, relative to the H37Rv reference genome, for all strain III isolates from the 'outbreak' village (i.e., 2007-2012). Each cluster is indicated (orange=IIIA, pink=IIIB, green=IIIC), with different shades within each cluster representing the different subgroups. Month/year of diagnosis are indicated for each isolate. SNPs defining clusters and subgroups are indicated in grey. Isolates identified for deep sequencing are indicated with arrows. From left to right, these are: MT-4942, 73787, MT-2184 and MT-504.

Isolates were selected for deep sequencing based on this site-by-site inspection and/or epidemiological significance and are indicated in **Table 5-2**. All other isolates had either unanimous base calls or were supported by >95% of reads.

TABLE 5-2 Allelic frequency for SNP loci defining clusters or subgroups by standard sequencing

H37Rv position	Gene name	Reference allele	Alternative allele	MT-4942	73787	MT-2184	MT-504
200530	*mce1B*	C	T	0,53	72,2	45,0	64,0
276685	*fadE4*	A	C	73,0	74,0	69,3	5,70
396246	*Rv0331*	A	C	48,5	81,1	59,1	70,1
921390	*Rv0828c*	T	C	87,0	70,0	53,0	77,0
1558108	*carB*	C	T	67,0	1,71	0,69	0,60
2082436	*Rv1835c*	C	T	62,1	68,0	53,0	88,0
3644579	*Rv3263*	G	A	96,0	0,68	0,66	0,58

The number of reads with the reference and alternative alleles for each position are indicated (ref,alt). 73787 selected for deep sequencing based on position at node in **Figure 5-5**.

Considering the above, it was possible that MT-4942, a smear positive case with strain IIIA diagnosed in 2007, exhibited micro-evolution and harboured bacteria with both the reference and alternative allele in *Rv0331* (48 reads were the reference allele and 5 reads were the alternative allele). Three individuals in cluster IIIA had the alternative allele in *Rv0331*. One (MT-5337) was diagnosed in 2010 and had an additional SNP in *Rv0630*. MT-5531 had TST conversion in early 2008, and subsequently developed disease in 2011, transmitting to MT-5983 within the same household. As MT-5531 lacked the SNP in *Rv0630*, this suggests that MT-4942 – who also lacked this SNP in *Rv0630* - was the original source of transmission. While it is possible that both MT-5337 and MT-5531 independently acquired SNPs in *Rv0331*, a more likely explanation is that MT-4942 carried bacteria with both reference and alternative alleles at this locus.

The above allelic frequencies suggested that micro-evolution was possible for MT-504 as well. This patient was diagnosed in December of 2011 with smear-positive cavitary disease, and represents the first case diagnosed in the n=5 subgroup of strain IIIB in **Figure 5-5**. While the majority of alleles (93%) support the alternative allele and this position was not called heterozygous, it is possible this patient initially had the reference allele, and subsequently acquired a SNP at this locus, which ultimately became the majority strain. In this scenario, MT-504 could have been transmitted to the n=13 subgroup of IIIB as well as the n=5 subgroup.

As alternative alleles in *fadE4* were noted for MT-2184 who was a smear-negative case diagnosed in 2012 (IIIB, n=13 subgroup in **Figure 5-5**), this isolate was also subjected to deep sequencing. 73787 (IIIB) was diagnosed in 2010 and selected based on its position (**Figure 5-5)**.

TABLE 5-3 Allelic frequency for SNP loci defining clusters or subgroups by deep sequencing

H37Rv position	Gene name	Reference allele	Alternative allele	MT-4942	73787	MT-2184	MT-504
200530	*mce1B*	C	T	0,281	177,0	193,1	151,0
276685	*fadE4*	A	C	246,0	193,1	177,5	8,144
396246	*Rv0331*	A	C	290,6	157,0	238,0	164,0
921390	*Rv0828c*	T	C	282,0	151,0	214,0	165,1
1558108	*carB*	C	T	277,0	0,161	0,229	0,169
2082436	*Rv1835c*	C	T	293,0	153,0	223,0	170,0
3644579	*Rv3263*	G	A	265,0	0,148	0,206	0,162

The number of reads with the reference and alternative alleles for each position are indicated (ref,alt). 73787 selected for deep sequencing based on position at node in **Figure 5-5**.

Overall, deep sequencing results do not provide strong evidence to support within-host heterogeneity, either due to mixed infection from >1 strain subgroup or micro-evolution. MT-4942 has 6 alternative alleles out of 296 reads (2%), while MT-504 has 8 reference alleles out of 152 reads (5%) at *fadE4*. While this is more indicative of sequencing or alignment error, it is not conclusive. Therefore, as epidemiologic data also support the possibility of micro-evolution, this cannot at present be ruled out.

CHAPTER 6. OBJECTIVE 2 – Manuscript II

Lee RS, Radomski N, Proulx J-F, Levade I, Shapiro BJ, McIntosh F, Soualhine H, Menzies D, Behr MA. Population genomics of *Mycobacterium tuberculosis* in the Inuit. *Proc Natl Acad Sci USA* 2015;112(44):13609-13614

6.1 Preamble

This Objective is an extension of our investigation of the 'outbreak' in the North. In the preceding decade, the maximum number of confirmed cases experienced by this community in a single year was 7, compared to 50 between 2011-2012. Coupled with the extraordinarily high attack rate among those with recent infection, this led to the concern that a new, hyper-virulent strain of *M. tuberculosis* had arrived in this village. These apprehensions were heightened by the occurrence of another apparent outbreak of similar scale in different Inuit community the following year.

To address these concerns, we conducted a population-based study of TB transmission in Nunavik over 23 years. The following manuscript describes the results of this study, and sheds light on the origins and epidemiology of TB in Nunavik.

The published reprint of this manuscript is enclosed in **Appendix 3-1**. This is followed by the accompanying supplementary data, including detailed methods, in **Appendix 3-2**. Supplementary datasets, for reader interest, are available open-access at http://www.pnas.org/content/112/44/13609.abstract?tab=ds

6.2 Manuscript II

Classification: Major category: Biological Sciences, Minor category: Evolution.

Population genomics of *Mycobacterium tuberculosis* in the Inuit

Robyn S. Lee [a,b,c,1], Nicolas Radomski [b,c,1], Jean-Francois Proulx [d], Ines Levade [e], B. Jesse Shapiro [e], Fiona McIntosh [b,c], Hafid Soualhine [f], Dick Menzies [b,c,g], Marcel A. Behr [b,c,2]

1. Equal contribution
2. Corresponding author

Author affiliations:

a. McGill University, Department of Epidemiology, Biostatistics and Occupational Health
b. The Research Institute of the McGill University Health Centre
c. McGill International TB Centre
d. Nunavik Regional Board of Health and Social Services
e. Département de sciences biologiques, Université de Montréal
f. Laboratoire de Santé Publique du Québec
g. Montreal Chest Institute, Respiratory Epidemiology and Clinical Research Unit

Corresponding author:

Dr. Marcel A. Behr
McGill International TB Centre
1001 boul Décarie, Block E, Mail Drop Point #EM33211,
Montréal, QC H4A 3J1 Canada
Phone: 514 934-1934, marcel.behr@mcgill.ca

Key words: *Mycobacterium tuberculosis*, evolution, whole genome sequencing.

Short title: Population genomics of *Mycobacterium tuberculosis*

Abstract

Nunavik, Québec suffers from epidemic tuberculosis (TB), with an incidence 50-fold higher than the Canadian average. Molecular studies in this region have documented limited bacterial genetic diversity among *Mycobacterium tuberculosis* isolates, consistent with a founder strain and/or ongoing spread. We have used whole-genome sequencing on 163 *M. tuberculosis* isolates from 11 geographically isolated villages to provide a high-resolution portrait of bacterial genetic diversity in this setting. All isolates were lineage 4 (Euro-American), with two sublineages present (major, $n = 153$; minor, $n = 10$). Among major sublineage isolates, there was a median of 46 pairwise single-nucleotide polymorphisms (SNPs), and the most recent common ancestor (MRCA) was in the early 20th century. Pairs of isolates within a village had significantly fewer SNPs than pairs from different villages (median: 6 vs. 47, $P < 0.00005$), indicating that most transmission occurs within villages. There was an excess of nonsynonymous SNPs after the diversification of *M. tuberculosis* within Nunavik: The ratio of nonsynonymous to synonymous substitution rates (dN/dS) was 0.534 before the MRCA but 0.777 subsequently ($P = 0.010$). Nonsynonymous SNPs were detected across all gene categories, arguing against positive selection and toward genetic drift with relaxation of purifying selection. Supporting the latter possibility, 28 genes were partially or completely deleted since the MRCA, including genes previously reported to be essential for *M. tuberculosis* growth. Our findings indicate that the epidemiologic success of *M. tuberculosis* in this region is more likely due to an environment conducive to TB transmission than a particularly well-adapted strain.

Introduction

The tubercule bacillus, *Mycobacterium tuberculosis*, is a highly successful, medically important human-adapted pathogen. Studies of diverse strain collections reveal a geographic aggregation of the principal *M. tuberculosis* lineages (1) consistent with a dissemination of this organism around the world with the paleo-migration (2). Ancient DNA studies also support the notion that *M. tuberculosis* has caused disease in humans for thousands of years. Thus, it can be inferred that *M. tuberculosis* has evolved in step with its human host, successfully responding to changes in the host and its environment that could affect the capacity to cause transmissible disease.

In contrast to the global diversity of *M. tuberculosis* strains (1–3), we have previously observed limited genetic diversity in the Nunavik region of Québec (4). One possible explanation is a founder strain, wherein genetic similarity is due to a single recent introduction of a bacterium and may not necessarily represent ongoing spread between communities. In this scenario, isolates might have indistinguishable genotypes by conventional genotyping modalities (restriction fragment length polymorphism, mycobacterial interspersed repetitive units, spoligotyping) but distinct genotypes when assessed using a higher-resolution method, namely whole-genome sequencing (WGS) (5). An additional explanation is that a single clone of *M. tuberculosis* is currently spreading both within and between villages; however, the great distances between these communities that are not linked by roads make intervillage spread less likely. These possible explanations need not be mutually exclusive.

To evaluate these possibilities, we conducted WGS on *M. tuberculosis* isolates from Nunavik isolated over 23 y. Estimation of the divergence date of the most recent common ancestor (MRCA) provided evidence that tuberculosis (TB) was introduced into this region in the early 20th century, following which time there has been substantial ongoing transmission, predominantly within villages. This setting provides a unique opportunity to study the genomic characteristics of an epidemiologically successful strain of *M. tuberculosis* over time.

Results

Whole-Genome Sequencing and Lineage Identification

There were 149 microbiologically confirmed TB cases diagnosed in Nunavik between 2001 and

2013; we obtained high-quality WGS data for 137/149 (92%). An additional 26 genomes were successfully sequenced from strains previously sampled between 1990 and 2000 (4). In total, WGS was conducted on 163 *M. tuberculosis* isolates. The average depth of coverage was 44.6× across 99.6% of the H37Rv reference genome.

All 163 genomes from the Nunavik region presented the 7-bp deletion in polyketide synthase (*pks*) *15/1* that characterizes lineage 4 of *M. tuberculosis* (the Euro-American lineage) (6). By comparing the Nunavik isolates with three genomes from each of the *M. tuberculosis* lineages (1–7), we observed that the 163 genomes were tightly clustered in two distinct sublineages: one consisting of 153 isolates (major; Mj) and the other consisting of 10 isolates (minor; Mn) (**Fig. 6-1**). Phylogenetic analyses based on single-nucleotide polymorphisms (SNPs) (**Figs. 6-1** and **6-2**) were supported by deletions confirmed by PCR (**Fig. S1** and Dataset S1).

Excluding SNPs in PE/PGRS and PPE genes, as well as mobile elements, as these may be at higher risk of false positives (5, 7), 1,288 single-nucleotide polymorphic loci were included comparing all genomes together against H37Rv (Dataset S2). The 153 isolates of the Mj sublineage had an average of 674 SNPs compared with H37Rv; the 10 isolates of the Mn sublineage had an average of 451 SNPs. There were 442 SNP loci shared across all Mj isolates, unique to this sublineage, and 214 SNP loci shared by all 10 Mn isolates that were not present in the Mj sublineage. According to the barcode proposed by Coll et al. (8) and the PhyTb tool of the PhyloTrack library (pathogenseq.lshtm.ac.uk/phytblive/index.php), the Mj and Mn sublineages can be classified as *M. tuberculosis* 4.1.2 and 4.8, respectively.

Phylogenetic analysis and the geographic distribution of isolates further distinguished the Mj and Mn sublineages (**Fig. 6-2**). To quantify diversity, we determined the number of pairwise SNPs within each sublineage. Among isolates of the Mj sublineage, the median number of pairwise SNPs was 46 [interquartile range (IQR) 13–49], with a maximum of 72. For isolates of the Mn sublineage, the median number of pairwise SNPs was 1 (IQR 0–2), with a maximum of 22. Nine of the 10 isolates from this sublineage were from the same village.

Transmission Occurs Mostly Within Villages.

To evaluate where most ongoing transmission occurs, we examined pairwise SNPs between isolates of the Mj sublineage, within and between villages, as these comprised over 90% of the cases in this region. The median number of pairwise SNPs was significantly lower for intravillage pairs (6, IQR 3–46) than for intervillage pairs (47, IQR 44–50, Wilcoxon–Mann–Whitney, P <0.00005). For both intra- and intervillage comparisons, a bimodal distribution was evident (**Fig. 6-3**). For the intravillage pairs (n = 3,689), the first mode comprised 61% of all pairwise comparisons and had a median of 3 SNPs (IQR 2–5). For the intervillage pairs (n = 7,939), the first mode comprised only 12% of all pairwise comparisons, and had a median of 9 SNPs (IQR 6–13). For both intra- and intervillage pairs, the second mode had similar distributions (median 47, IQR 45–49 and median 48, IQR 45–50, respectively), consistent with the star-like pattern shown in **Figs. 6-1** and **6-2**.

We also considered thresholds for transmission based on published *M. tuberculosis* substitution rates (0.5 SNPs per genome per y, 95% confidence interval 0.3–0.7) (5, 9). For a study spanning 23 y (1991–2013 inclusive), we expected that epidemiologically linked cases would be separated by no more than 12 SNPs. Applying this threshold, 2,208 of 3,689 (60%) intravillage pairs were separated by 12 or fewer SNPs, compared with 683 of 7,939 (9%) intervillage pairs (two-sample z test for difference in proportions, P <0.00005). Sensitivity analyses applying substitution rates of 0.3 and 0.7 SNPs per genome per y yielded similar results.

***M. tuberculosis* diversified in Nunavik during the 20th Century**

Relative to the global genetic diversity of *M. tuberculosis*, the total diversity of strains in Nunavik was low, consistent with a recent introduction of TB into this region. To evaluate this hypothesis, we estimated the MRCAs for each sublineage using Bayesian molecular dating (10, 11). Constraining the substitution rate of *M. tuberculosis* based on previous estimates (5, 9) we inferred the MRCA of the Mj sublineage to be 1919 [95% highest posterior density interval (HPD) 1892–1946], with other divergence dates within the Mj sublineage scattered over the 20th century (**Table 6-1**, analysis 1). The Mn sublineage was found to have an MRCA of 1976 (95% HPD 1951–1994). Repeating these analyses without constraining the substitution rate yielded similar results (**Table 6-1**, analyses 2 and 3).

Natural selection of *M. tuberculosis* in a new environment

The *M. tuberculosis* population may have experienced a new regime of natural selection upon its introduction into Nunavik. First, *M. tuberculosis* could have experienced a population bottleneck upon introduction, reducing the efficacy of natural selection and allowing the fixation of deleterious mutations. Second, upon entry into a new environment, *M. tuberculosis* could have experienced positive selection, retaining fitter variants over time. Third, if the environment was conducive to transmission of *M. tuberculosis*, there may have been a relaxation of purifying selection across the entire genome. These scenarios are not mutually exclusive, and other scenarios are possible as well.

To measure natural selection at the protein level, we used the ratio of nonsynonymous to synonymous substitution rates (dN/dS), reasoning that this should remain stable over time in the absence of changing regimes of natural selection (12). Specifically, we tested the null hypothesis that dN/dS remained the same pre- and postdiversification of each *M. tuberculosis* sublineage. We first reconstructed the ancestral sequences of the MRCA for each sublineage, along with that of the common ancestor for these two sublineages (denoted "Mj–Mn"). We then compared the nonsynonymous and synonymous SNPs (nsSNPs and sSNPs, respectively) between these reconstructed ancestors (Mj–Mn versus Mj, Mj–Mn versus Mn) to obtain the dN/dS for each sublineage prediversification. To calculate the dN/dS postdiversification (i.e., subsequent to the MRCAs for each sublineage), we generated a concatenated sequence of codons for both the Mj and Mn sublineages that included all SNP loci and compared each sequence with that of its respective ancestor. In this phylogenetic approach, each independent SNP was counted exactly once (i.e., SNPs present in multiple isolates were not recounted). In total, for the Mj sublineage, we identified 229 nsSNPs and 154 sSNPs before its introduction into Nunavik, compared with 238 nsSNPs and 107 sSNPs that occurred subsequently (Dataset S2). The dN/dS ratio for SNPs prediversification was 0.534, consistent with published estimates for *M. tuberculosis* (13), whereas the dN/dS postdiversification was 0.777 (**Table 6-2**, analysis 1a; *G* test based on numbers of nsSNPs and sSNPs pre- and postdiversification, $P = 0.010$). Singleton SNPs, present in only one isolate, are expected to be enriched in nonsynonymous mutations destined to be purged by purifying selection. To evaluate whether the increased dN/dS was attributable to these

transient mutations, we restricted our analysis to SNPs present in ≥2 isolates. We still observed a significant increase in the dN/dS, going from 0.534 prediversification to 0.928 postdiversification (**Table 6-2**, analysis 1b). There was no significant difference in postdiversification nsSNPs and sSNPs comparing analyses with and without singletons (Fisher's exact test, $P = 0.472$). As an alternative method of calculating the dN/dS postdiversification, we conducted a pairwise analysis wherein the median dN/dS was obtained by comparing each of the 153 Mj isolates with its respective ancestral sequence. This yielded similar results, whether singletons were included or excluded (**Table 6-2**, analysis 2). Compared with the Mj sublineage, the dN/dS ratios for the Mn sublineage were more stable over time (**Table 6-2**).

The efficiency of purifying selection to remove deleterious nonsynonymous mutations is reduced when populations undergo dramatic size fluctuations due, for example, to bottlenecks or exponential growth. To investigate whether the increased dN/dS ratio in the Mj sublineage was due to an expanding bacterial population size over time, we constructed Bayesian skyline plots (**Fig. S2**). Model comparison using Akaike's information criterion for Markov chain Monte Carlo samples [AICM (14)] rejected an exponential population growth in favor of a constant population size or Bayesian skyline model (**Table S1**). Together, these results suggest that the genome-wide increase in dN/dS was not due to a population bottleneck followed by exponential growth, nor to a lack of time for purifying selection to purge deleterious nsSNPs.

Genes affected by SNPs and/or deletions

Unlike genome-wide relaxation, wherein the whole genome is affected, positive selection is thought to target specific genes (15, 16). Across the 153 genomes of the Mj sublineage, we identified 218 and 227 genes with nsSNPs pre- and postdiversification, respectively (Dataset S2). To evaluate whether any particular categories of *M. tuberculosis* genes were unusually variable postdiversification, we tabulated these SNPs according to gene categories described in the literature (**Fig. 6-4** and Datasets S3 and S4). There was no statistically significant difference between the proportion of genes with nsSNPs in any categories pre- versus postdiversification (two-sample *z* test for difference in proportions, $P > 0.05$). However, genes predicted to be conditionally essential for *M. tuberculosis* survival in vitro, in macrophages, or in vivo were not spared nsSNPs (Dataset S5). Mutations in essential genes often affected a residue that is

conserved in the closely related mycobacterial species *Mycobacterium canettii* (17) and *Mycobacterium kansasii* (18), with three genes (*Rv0338c*, *echA5*, and *murC*) incurring distinct nsSNPs in different strains (Dataset S5).

In addition to these potentially deleterious SNPs, all Mj isolates lacked eight regions, resulting in 13 deleted genes. Certain strains also suffered a further seven deletions, disrupting 28 genes (Dataset S1). Certain gene categories appeared overrepresented in postdiversification deletions (e.g., genes acquired through lateral gene transfer, mobile elements), but the low number of deleted genes precluded robust statistical analysis (**Fig. 6-4** and Dataset S1). Four genes predicted to be essential in genomic screens were completely (*Rv2335*) or partially (*Rv1939*, *Rv2885c*, and *Rv3135*) deleted in some isolates of the Mj sublineage (Dataset S3). *Rv2335* (i.e., *cysE*) codes for a serine acetyltransferase, predicted to be essential for survival in vivo (19), that was absent in eight isolates. *Rv2885c* codes for a transposase in the IS*1539* insertion sequence that is predicted to be essential for survival in vivo (19), whereas *Rv3135* codes for *PPE50* and is predicted to be essential for survival in vitro (20). *Rv1939* codes for an oxydoreductase predicted to be essential for survival in vivo (19) that was deleted in one isolate (18421) (Dataset S1).

Discussion

The Inuit originally came from eastern Siberia, via the Bering Strait, in two waves over several thousands of years (21). Given the recognized close association between *M. tuberculosis* and human populations, it is theoretically possible that they brought an East Asian lineage of *M. tuberculosis* with them to the Canadian Arctic. Our data refute this scenario by revealing only lineage 4 (Euro-American) isolates. The low amount of genetic diversity among isolates from different villages indicates that the vast majority of TB cases in this region are the consequence of a single introduction of *M. tuberculosis*, perhaps from Europe, around the early 20th century. The introduction and diversification of a single dominant clone in Nunavik provide an unobstructed view of *M. tuberculosis* over time, enabling us to draw certain inferences about the epidemiology and evolution of this highly successful human-adapted pathogen.

The Inuit have had casual interactions with Europeans since the 17th century, most notably with whalers and explorers who sailed along the coasts of Hudson's Bay and Labrador (22). However,

the first permanent settlements of the Hudson's Bay Company in the region now known as Nunavik date to the late 19th and early 20th centuries, following which there were more sustained interactions between the Inuit and traders (23). Our MRCA estimates support an introduction of TB into this region during this period, which is consistent with some, but not all, historical accounts of when TB was first observed (24). The apparent lack of TB before the early 20th century, despite several centuries of Inuit–European interactions, supports that TB is generally not spread through casual contact, as is the case for measles or chickenpox. This is also consistent with our analysis of the pairwise SNPs between isolates across villages; only a small proportion of intervillage case pairs had low SNP differences, arguing against transmission during casual contact, as can occur at cultural gatherings that bring together members of different villages. Supporting this, villages often had one predominant strain, and individual strains were mostly confined to one village (**Fig. 6-2**). This observation presents both an opportunity and a challenge for public health; whereas TB should in theory be amenable to control through scaled-up efforts, it may be that village-by-village, rather than regional, interventions will be needed to interrupt transmission in this setting.

In a number of high-incidence countries, the emergence of an epidemiologically successful strain has been attributed to virulence features encoded in the bacterial genome (25). For instance, the polyketide synthase-derived phenolic glycolipid (PGL) coded by the intact *pks15/1* locus of strain HN878 (Beijing genotype) induces hyperlethality in murine disease models (26), potentially explaining the emergence of the Beijing strain in a number of settings worldwide (27). Furthermore, compared with other clinical strains, strains 1471 and HN878 (Beijing genotypes) result in increased macrophage necrosis (28) and more progressive pathology in experimental infections (29). However, although certain strains have a propensity to cause accelerated life-threatening pathology in experimental models, it is not yet clear whether this property predicts epidemiologic success, as a strain that causes chronic, nonprogressive pathology may be the most likely to transmit.

In Nunavik, we observed a set of related strains that meet the epidemiologic criterion of success, without any clear genomic indicators of increased bacterial virulence. Instead, for the Mj sublineage, we observe an enrichment of nsSNPs since its introduction into this region, some of

which are expected to affect the function of proteins that contribute to the survival of *M. tuberculosis* during infection. There are a number of potential causes of an increased dN/dS, including insufficient time for purifying selection to act, positive selection, relaxed purifying selection, and genetic drift. Whereas an increased dN/dS at the tips of a phylogenetic tree may indicate insufficient time for purifying selection (13), the postdiversification inflation of dN/dS holds even with the exclusion of evolutionarily recent singleton SNPs. Therefore, a simple time dependence is unlikely to be the only explanation. Positive selection is unlikely to inflate the dN/dS across the entire genome but rather should target genes with specific functions (15, 16). Although we did not identify any particular functional category of genes enriched in nsSNPs, this does not exclude positive selection on a small number of genes. However, it suggests that positive selection was not the pervasive force leading to a high dN/dS genome-wide. The remaining potential explanations for the dN/dS elevation are a genome-wide relaxation of purifying selection and genetic drift. The nsSNPs and deletions in putatively essential genes provide further support for these two interpretations.

The global *M. tuberculosis* population has been previously shown to evolve through mostly weak selection and strong drift (30); here we show that the same is true on a local level, to an even greater extent. Given that drift will have stronger effects when effective populations are reduced (31) and that our data suggest that population size remained more or less constant, we hypothesize that relaxation of purifying selection has contributed significantly to the evolution of the Nunavik strain of *M. tuberculosis*. Further investigation in this and other similar populations is needed. Regardless of the forces that have driven the elevated dN/dS, our findings suggest that *M. tuberculosis* has not thrived in Nunavik due to a unique virulence profile of the bacteria. It follows that *M. tuberculosis* control in this region, and in similar settings, will require looking beyond the bacterial culprit to the social conditions that foster TB.

Materials and Methods

Detailed methods can be found in ***SI Materials and Methods***. In brief, the Nunavik region is composed of 14 Inuit communities, with a total population of 12,090 (in 2011). Between 1990 and 2013, there were 200 cases of TB in Nunavik, of which 163 were available for whole-genome sequencing using the MiSeq 250 System (Illumina). Reads were assembled and

compared as previously described (32). The final dataset of SNPs excluded those in PE/PGRS and PPE genes, as well as mobile elements, as these may be prone to false positives (5, 7). Deletion events were identified with the Integrative Genomics Viewer (33) and confirmed by PCR and Sanger sequencing. Concatenated sequences of the SNPs were used to generate phylogenetic trees via the maximum likelihood method in Molecular Evolutionary Genetics Analysis [MEGA (34)]. Divergence times for the 163 Nunavik isolates were estimated using Bayesian Markov chain Monte Carlo methods [Bayesian Evolutionary Analysis by Sampling Trees (10, 11)], with H37Rv used as an outgroup.

We used three approaches to derive MRCAs. Using the concatenated sequences of SNPs across the 163 genomes, we first conducted an analysis that incorporated prior knowledge of the substitution rate of *M. tuberculosis* in the form of a calibration node for the Mj sublineage (analysis 1). We then performed an analysis agnostic to the reported substitution rate (i.e., without calibration), also using concatenated sequences (analysis 2). We then repeated this second analysis but applied a correction for the constant sites across the genomes (analysis 3). Different coalescent models were tested to explore changes in effective population size over time (35). The AICM (14) was used to select the model providing the best fit. Bayesian skyline plots were generated (**Fig. S2**).

To calculate the dN/dS ratios, the ancestral sequences for each MRCA (Mj–Mn, Mj, and Mn) were reconstructed manually (Dataset S2). We then calculated the dN/dS pre- and postdiversification for the Mj and Mn sublineages, using both a phylogenetics-based approach (analysis 1) and a pairwise dN/dS analysis (analysis 2) (7). For both analyses, we repeated the dN/dS calculations after excluding SNPs that were present only once across all 163 genomes (singletons).

Ethical approval for this work was obtained from the McGill University Faculty of Medicine Institutional Review Board.

FIGURES

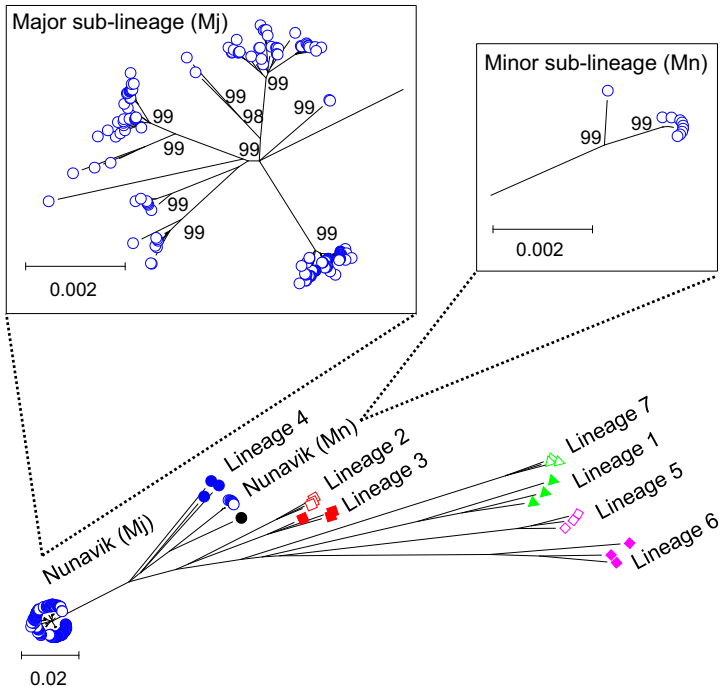


FIGURE 6-1. Maximum likelihood tree of 163 *M. tuberculosis* isolates from Nunavik and 21 representative genomes of lineages 1–7. Phylogenetic clusters based on 9,016 single-nucleotide polymorphic loci identified across 184 genomes compared with H37Rv (solid black circle). The scale bars represent the number of substitutions per site. Bootstrap values from 1,000 replicates are shown for branches within the Mj and Mn sublineages. For clarity, only values ≥98 are shown.

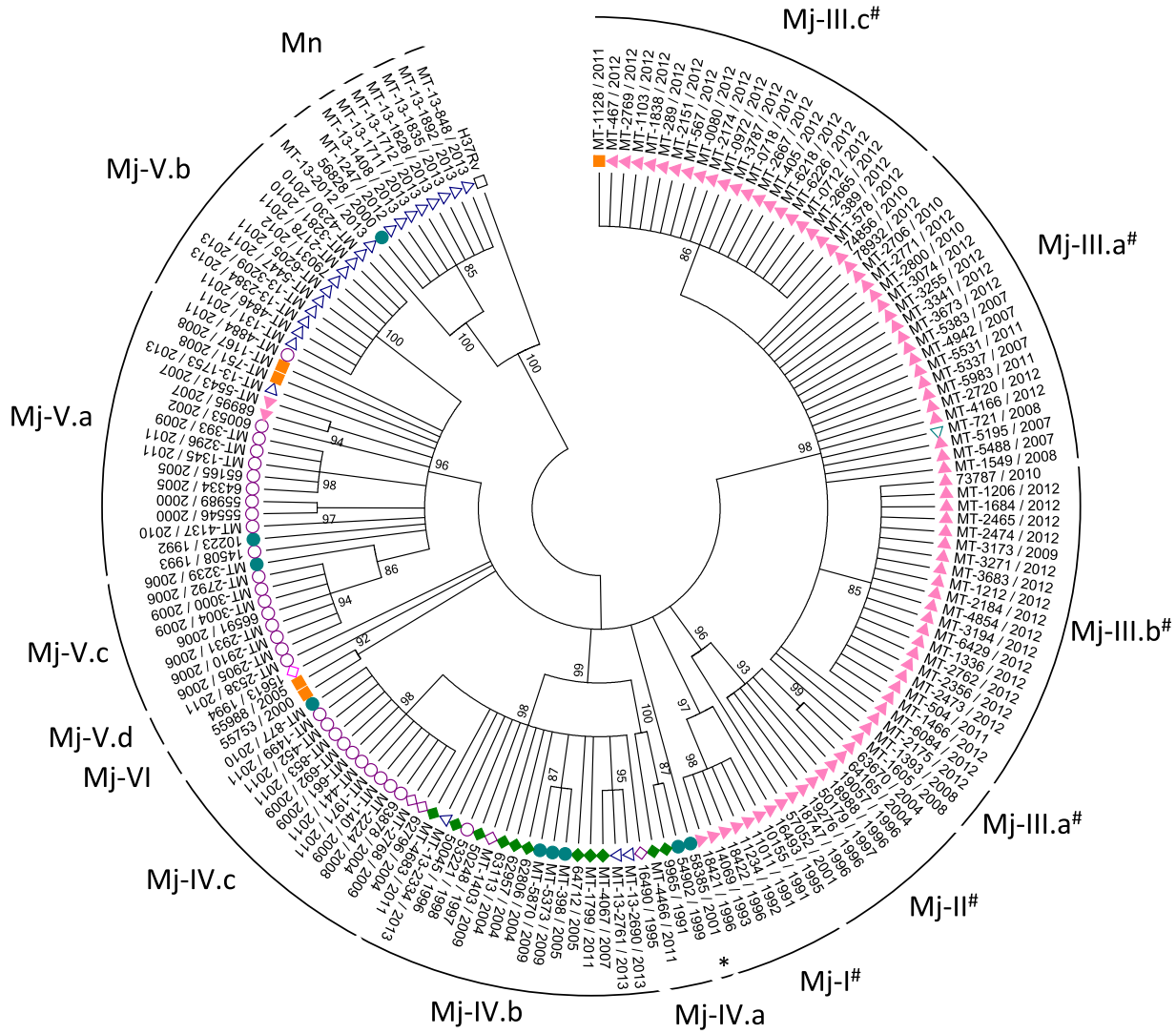


FIGURE 6-2. Maximum likelihood tree of 163 *M. tuberculosis* isolates from Nunavik. Phylogenetic clusters were identified based on 1,288 single-nucleotide polymorphic loci compared with H37Rv. Solid and dashed lines indicate isolates of the Mj and Mn sublineages, respectively. Colored shapes represent the reference genome (bordered black square) and the villages of Nunavik: A (bordered blue triangle), B (full orange square), C (bordered purple circle), D (full green diamond), E (bordered purple diamond), K (full pink triangle), and other (full green circle). *Genome with a unique single-nucleotide polymorphism profile.

#Phylogenetic clusters defined previously in ref. 32. Years of diagnosis are indicated. Bootstrap support from 1,000 replicates is shown. Branches supported by less than 80% of bootstrap replicates are collapsed.

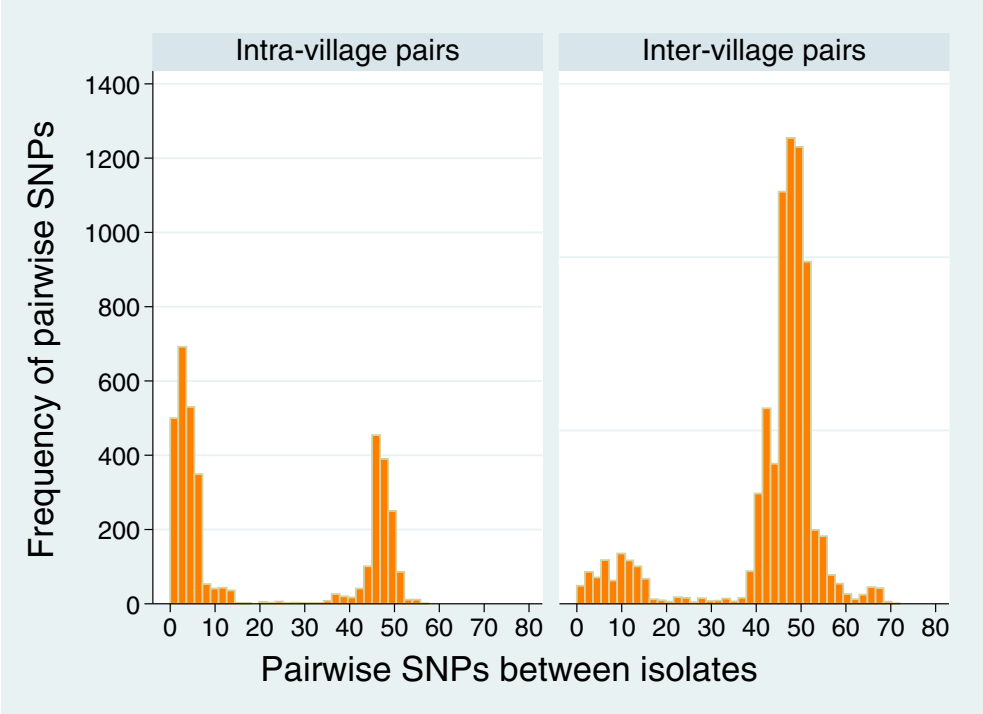


FIGURE 6-3. Pairwise SNPs between isolates of the major sublineage of Nunavik. There were a total of 11,628 pairwise comparisons: 3,689 intravillage case pairs and 7,939 intervillage case pairs.

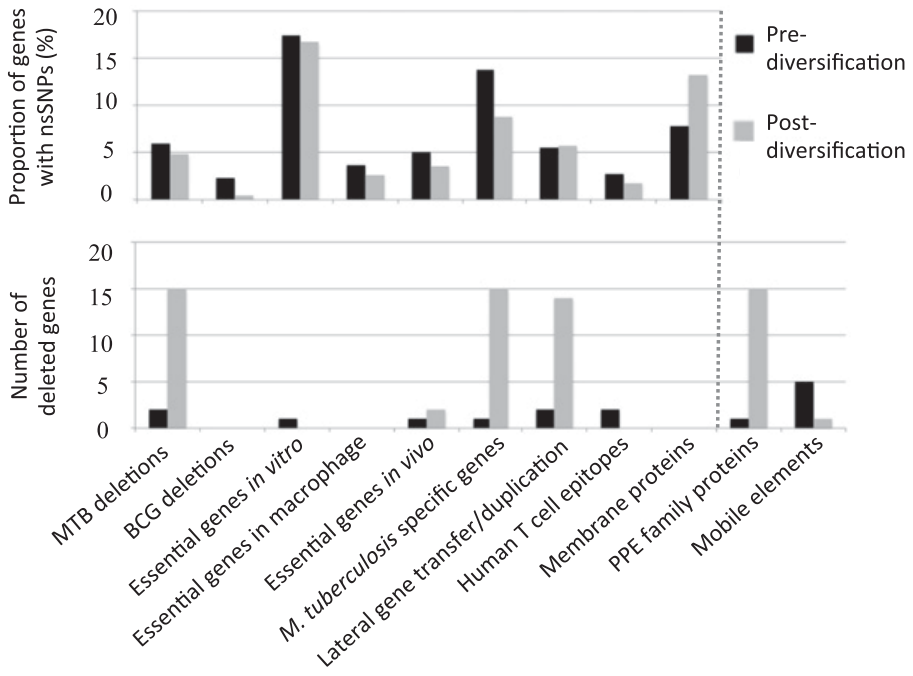


FIGURE 6-4. Proportion of genes with nonsynonymous single-nucleotide polymorphisms (Top) and the number of deleted genes (Bottom) for the major sublineage, pre- and post-diversification. Gene categories are as defined in the publications: *M. tuberculosis* (MTB) deletions (36), bacillus Calmette–Guérin (BCG) deletions (37), essential genes in vitro (20), in macrophages (38), or in vivo (19), *M. tuberculosis*-specific genes (39), lateral gene transfer or duplication acquisition (39), human T-cell epitopes (7), genes coding membrane proteins (40), mobile elements (7), and genes coding PPE family proteins (7). Genes designated as PE/PGRS, PPE, or mobile elements were excluded from the SNP analysis (7).

TABLES

TABLE 6-1. Estimated year of divergence of *M. tuberculosis* sublineages and clusters of Nunavik.

Phylogenetic sublineages and clusters	Analysis 1 *	Analysis 2	Analysis 3
Mj-Mn	1053 (602-1450)	1243 (836-1575)	744 (230-1216)
Mj	1919 (1892-1946)	1922 (1890-1950)	1904 (1873-1930)
Mj-I-II[#]	1942 (1919-1964)	1947 (1921-1967)	1925 (1898-1948)
Mj-V	1952 (1929-1973)	1956 (1932-1978)	1935 (1909-1958)
Mj-IV	1965 (1949-1978)	1966 (1951-1980)	1958 (1941-1973)
Mj-III.a.b.c[#]	1999 (1993-2004)	1999 (1993-2004)	2000 (1993-2004)
Mj-VI	1999 (1995-2000)	1999 (1995-2000)	1999 (1995-2000)
Mn	1976 (1951-1994)	1979 (1953-1997)	1969 (1943-1987)

All numbers expressed in calendar years, rounded to the nearest whole number. Analysis 1: Calibration point, concatenated alleles. Analysis 2: No calibration point, concatenated alleles. Analysis 3: No calibration point, weighting for constant sites. The median date of divergence is shown in years, with corresponding 95% highest posterior density intervals. * Results of this analysis are reported in text. [#] Strain code as per Lee *et al*. 2015 (34).

TABLE 6-2. dN/dS of *M. tuberculosis* sublineages pre and post-diversification in Nunavik.

	Mj sublineage			Mn sublineage		
Analysis	**Pre-diversification**	**Post-diversification**	**P value**	**Pre-diversification**	**Post-diversification**	**P value**
1a – *all SNPs*	0.534	0.777	0.010	0.547	0.615	0.873
1b – *excluding singletons*	0.534	0.928	0.005	0.547	0.759	0.767*
2a – *all SNPs*	0.534	0.947	<0.00005	0.547	0.759	0.006
2b – *excluding singletons*	0.534	0.953	<0.00005	0.547	0.759	0.006

Ancestral sequences were reconstructed for the MRCA of the Mj-Mn sublineages, as well as the Mj sublineage and the Mn sublineage. Pre-diversification: 229 nonsynonymous (ns)SNPs and 154 synonymous (s)SNPs identified in the Mj sublineage, and 113 nsSNPs and 75 sSNPs in the Mn sublineage. Analysis 1a: The dN/dS pre-diversification was calculated by comparing ancestral sequences. For post-diversification, concatenated sequences of codons for each sublineage were generated based on all SNP loci identified, with SNPs in more than isolate only contributing once. Overall, there were 238 nsSNPS and 107 sSNPs in the Mj sublineage and 13 nsSNPs and 8 sSNPs in the Mn. These concatenated sequences were then compared to their respective ancestral sequences to obtain a dN/dS. The raw counts of non-redundant nsSNPs and sSNPs pre- and post-diversification were compared for each sublineage using the G-test, with p values shown. Analysis 1b: Excluding singleton SNPs. G-test based on 120 nsSNPs and 46 sSNPs for Mj and 8 nsSNPs and 4 sSNPs for Mn post-diversification. Analysis 2a: dN/dS was calculated for each isolate compared to its respective ancestral sequence (i.e. 153 Mj isolates were compared to the imputed ancestral sequence for Mj). Within each sublineage, the median dN/dS was calculated and is shown above. Analysis 2b: Excluding singleton SNPs. The Wilcoxon Signed Rank Test was used to compare the median dN/dS post-diversification for each sublineage with its respective pre-diversification estimate. *Fisher's Exact Test due to cell counts <5.

References

1. Gagneux S, Small PM. (2007) Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development. Lancet Infect Dis 7(5):328–337.
2. Wirth T, *et al*. (2008) Origin, spread and demography of the Mycobacterium tuberculosis complex. PLoS Pathog 4(9):e1000160.
3. Gagneux S, *et al*. (2006) Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. Proc Natl Acad Sci USA 103(8):2869–2873.
4. Nguyen D, *et al*. (2003) Tuberculosis in the Inuit community of Quebec, Canada. Am J Respir Crit Care Med 168(11):1353–1357.
5. Roetzer A, *et al*. (2013) Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: A longitudinal molecular epidemiological study. PLoS Med 10(2):e1001387.
6. Marmiesse M, *et al*. (2004) 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(Pt 2):483–496.
7. Comas I, *et al*. (2010) Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. Nat Genet 42(6):498–503.
8. Coll F, *et al*. (2014) A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. Nat Commun 5:4812.
9. Walker TM, *et al*. (2013) Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: A retrospective observational study. Lancet Infect Dis 13(2):137–146.
10. Bouckaert R, *et al*. (2014) BEAST 2: A software platform for Bayesian evolutionary analysis. PLOS Comput Biol 10(4):e1003537.
11. Drummond AJ, Suchard MA, Xie D, Rambaut A. (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29(8):1969–1973.
12. McDonald JH, Kreitman M. (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature 351(6328):652–654.
13. Rocha EPC, *et al*. (2006) Comparisons of dN/dS are time dependent for closely related bacterial genomes. J Theor Biol 239(2):226–235.
14. Baele G, *et al*. (2012) Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. Mol Biol Evol 29(9):2157–

2167.
15. Novichkov PS, Wolf YI, Dubchak I, Koonin EV. (2009) Trends in prokaryotic evolution revealed by comparison of closely related bacterial and archaeal genomes. J Bacteriol 191(1):65–73.
16. Shapiro BJ, Alm EJ. (2008) Comparing patterns of natural selection across species using selective signatures. PLoS Genet 4(2):e23.
17. Supply P, *et al*. (2013) Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of *Mycobacterium tuberculosis*. Nat Genet 45(2):172–179.
18. Wang J, *et al*. (2015) Insights on the emergence of *Mycobacterium tuberculosis* from the analysis of *Mycobacterium kansasii*. Genome Biol Evol 7(3):856–870.
19. Sassetti CM, Rubin EJ. (2003) Genetic requirements for mycobacterial survival during infection. Proc Natl Acad Sci USA 100(22):12989–12994.
20. Sassetti CM, Boyd DH, Rubin EJ (2003) Genes required for mycobacterial growth defined by high density mutagenesis. Mol Microbiol 48(1):77–84.
21. Raghavan M, *et al*. (2014) The genetic prehistory of the New World Arctic. Science 345(6200):1255832.
22. Higdon J. (2010) Commercial and subsistence harvests of bowhead whales (*Balaena mysticetus*) in eastern Canada and west Greenland. J Cetacean Res Manag 11:185.
23. Bonesteel S. (2006) Canada's Relationship with the Inuit, ed Anderson E (published under the authority of the Minister of Indian Affairs and Northern Development and Federal Interlocutor for Métis and Non-Status Indians, Ottawa, Canada).
24. Grygier PS (1994) A Long Way from Home: The Tuberculosis Epidemic Among the Inuit (McGill-Queen's Univ Press, Montreal).
25. Alonso H, *et al*. (2011) Deciphering the role of IS6110 in a highly transmissible *Mycobacterium tuberculosis* Beijing strain, GC1237. Tuberculosis (Edinb) 91(2):117–126.
26. Reed MB, *et al*. (2004) A glycolipid of hypervirulent tuberculosis strains that inhibits the innate immune response. Nature 431(7004):84–87.
27. Parwati I, van Crevel R, van Soolingen D. (2010) Possible underlying mechanisms for successful emergence of the *Mycobacterium tuberculosis* Beijing genotype strains. Lancet Infect Dis 10(2):103–111.
28. Amaral EP, *et al*. (2014) Pulmonary infection with hypervirulent Mycobacteria reveals a

crucial role for the P2X7 receptor in aggressive forms of tuberculosis. PLoS Pathog 10(7):e1004188.

29. Ordway D, *et al*. (2007) The hypervirulent *Mycobacterium tuberculosis* strain HN878 induces a potent TH1 response followed by rapid down-regulation. J Immunol 179(1):522–531. 30.
30. Hershberg R, *et al*. (2008) High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. PLoS Biol 6(12):e311. 31.
31. Kuo CH, Moran NA, Ochman H. (2009) The consequences of genetic drift for bacterial genome complexity. Genome Res 19(8):1450–1454. 32.
32. Lee RS, *et al*. (2015) Reemergence and amplification of tuberculosis in the Canadian Arctic. J Infect Dis 211(12):1905–1914. 33.
33. Thorvaldsdóttir H, Robinson JT, MesirovJP (2013). Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Brief Bioinform 14(2):178–192. 34.
34. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol 30(12):2725–2729. 35.
35. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. (2006) Relaxed phylogenetics and dating with confidence. PLoS Biol 4(5):e88. 36.
36. Tsolaki AG, *et al*. (2004) Functional and evolutionary genomics of *Mycobacterium tuberculosis*: Insights from genomic deletions in 100 strains. Proc Natl Acad Sci USA 101(14):4865–4870. 37.
37. Mostowy S, Tsolaki AG, Small PM, Behr MA. (2003) The in vitro evolution of BCG vaccines. Vaccine 21(27–30):4270–4274. 38.
38. Rengarajan J, Bloom BR, Rubin EJ. (2005) Genome-wide requirements for *Mycobacterium tuberculosis* adaptation and survival in macrophages. Proc Natl Acad Sci USA 102(23):8327–8332. 39.
39. Stinear TP, *et al*. (2008) Insights from the complete genome sequence of Mycobacterium marinum on the evolution of *Mycobacterium tuberculosis*. Genome Res 18(5):729–741.
40. Osório NS, *et al*. (2013) Evidence for diversifying selection in a set of *Mycobacterium tuberculosis* genes in response to antibiotic- and nonantibiotic-related pressure. Mol Biol Evol 30(6):1326–1336. 41.

41. Cingolani P, *et al*. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w^{1118}; iso-2; iso-3. Fly 6(2):80–92. 42.
42. Rutherford K, *et al*. (2000) Artemis: Sequence visualization and annotation. Bioinformatics 16(10):944–945. 43.
43. Waddell PJ, Steel MA. (1997) General time-reversible distances with unequal rates across sites: Mixing gamma and inverse Gaussian distributions with invariant sites. Mol Phylogenet Evol 8(3):398–414. 44.
44. Tamura K. (1992) Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. Mol Biol Evol 9(4):678–687. 45.
45. Saitou N, Nei M. (1987) The Neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol Biol Evol 4(4):406–425. 46.
46. Felsenstein J. (1985) Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39(4):783–791. 47.
47. Comas I, *et al*. (2013) Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. Nat Genet 45(10):1176–1182. 48.
48. Steenken W, Oatway WH, Petroff SA. (1934) Biological studies of the tubercle bacillus: III. Dissociation and pathogenicity of the R and S variants of the human tubercle bacillus (H_{37}). J Exp Med 60(4):515–540. 49.
49. Rambaut A, Suchard M, Xie D, Drummond AJ. (2014) Tracer v1.6. Available at beast.bio.ed.ac.uk/software/tracer.
50. Nei M, Gojobori T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3(5):418–426.

6.3 Additional analyses

The tMRCAs presented for each sublineage in **Table 6-2** were extracted from the following time tree, which has been converted to time in years using a calibration node and dates of collection for each of the isolates (analysis 1).

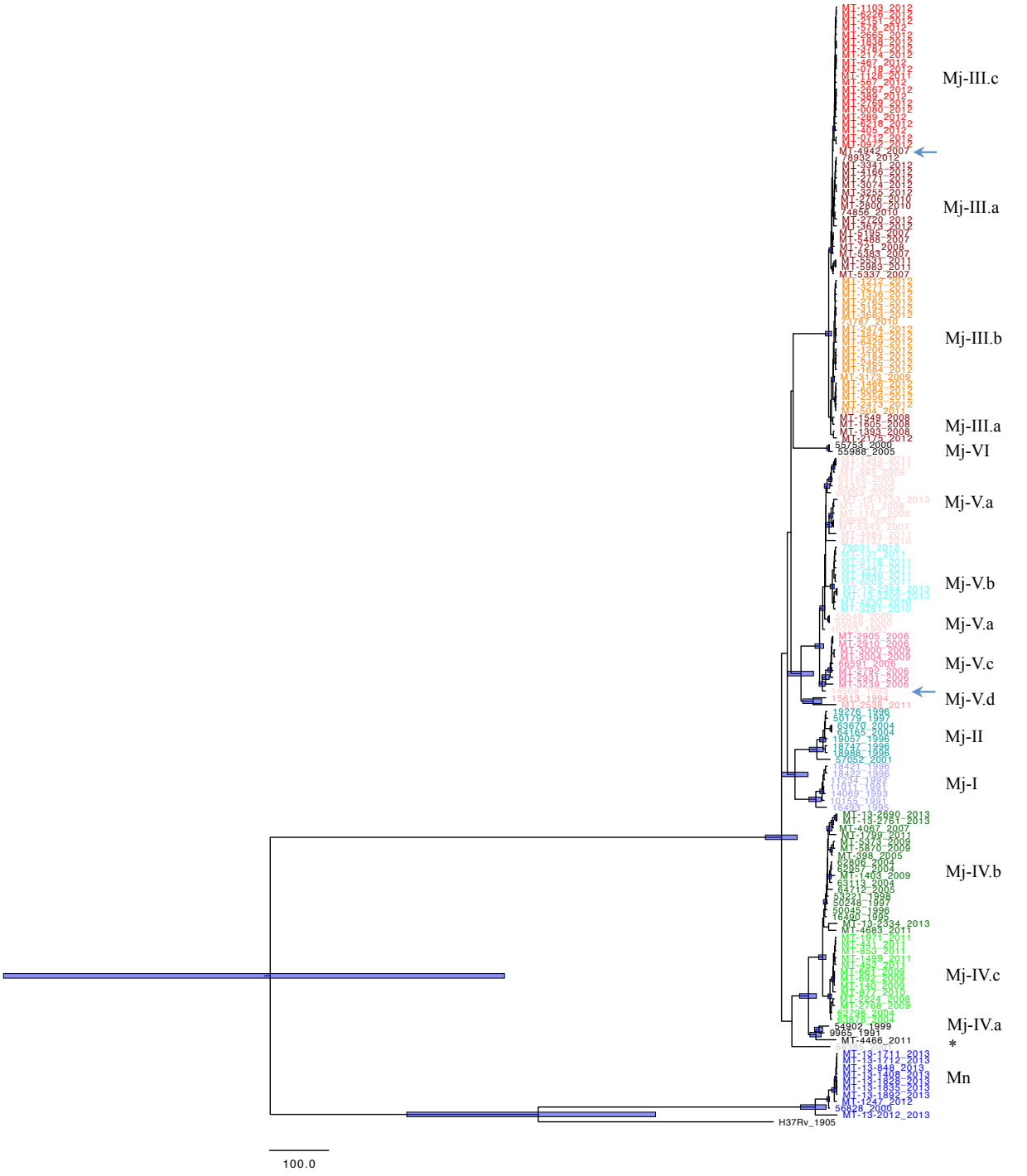


FIGURE 6-5. Maximum clade credibility tree of 163 *M. tuberculosis* isolates from Nunavik. This tree was produced using Bayesian Evolutionary Analysis by Sampling Trees (175), as described in the Methods of Manuscript II (analysis 1). All isolates are coloured by cluster, based on cluster assignments from the main manuscript using the maximum likelihood (ML) method. While some variation is evident compared to the ML tree, isolates remained with the same clusters. Two exceptions are MT-4942, indicated near the top with an arrow; this isolate moved from Mj-III.a to the an ancestral position in Mj-III.c despite having 0 SNPs from many in IIIA and a minimum of 2 SNPs difference compared to any III.a isolate. Isolate 14508 also moved from the closely-related Mj-V.a to Mj-V.c (also indicated with an arrow). 95% highest posterior density intervals are shown at nodes that had posterior density >0.8, i.e., these nodes (and isolates contained therein) were present in at least 80% of the sampled trees. A scale in calendar years is indicated.

CHAPTER 7. OBJECTIVE 3 – Manuscript III

Lee RS, Proulx J-F, Menzies D, Behr MA. Progression to tuberculosis disease increases with multiple exposures. Under review at *Eur Respir J.*

7.1 Preamble

This Objective represents the final investigation of the 2011-2012 'outbreak' in Nunavik. In the previous manuscript, analyses did not support the recent introduction of a hyper-virulent strain in this region. Two case-control studies conducted in this village in 2013 suggested housing occupancy might play a role in progression to disease; however, this was only among those residing with smear positive cases, of which there were few. Thus, we had not identified a potential risk factor that could account for the elevated attack rate in this community. Given the observation that many cases had multiple contacts with other cases during the 'outbreak', we considered an alternative hypothesis: that multiple exposures were associated with increased odds of progression.

The accompanying supplemental data, which includes detailed methods and additional analyses, can be found in **Appendix 4**.

7.2 Manuscript III

Progression to Tuberculosis Disease Increases with Multiple Exposures

Robyn S. Lee[1,2,3], Jean-François Proulx[5], Dick Menzies[2,3,6], Marcel A. Behr[1-4*]

1. Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada
2. McGill International TB Centre, Montreal, Quebec, Canada
3. The Research Institute of the McGill University Health Centre, Montreal, Quebec, Canada
4. Department of Medicine, McGill University, Montreal, Quebec, Canada
5. Nunavik Regional Board of Health and Social Services, Kuujjuaq, Quebec, Canada
6. Respiratory Epidemiology and Clinical Research Unit, Montreal Chest Institute, Montreal, Quebec, Canada

* Corresponding author:
Marcel A. Behr
McGill University Health Centre
1001 boul Décarie
Mail Drop Point #E05.1115
Montréal, QC H4A 3J1 Canada
Phone: 514 934-1934 (42815)
Fax: (514) 934-4476
email: marcel.behr@mcgill.ca

Running head: Multiple Exposures
Word count: 2777

Take home message: Increased exposure may not only be associated with risk of infection, but also risk of progression to TB disease.

Abstract

During a single year, a Canadian village had 34 individuals with microbiologically-confirmed TB among 169 with new infection (20%). Contact investigation revealed multiple exposures for each person. We investigated whether intensity of exposure might contribute to this extraordinary risk of disease.

Materials and methods Case-control study. Among those with new infection, 34 had culture-confirmed TB (cases) and 118 did not progress (controls), excluding 17 with probable disease. Contact investigation data were utilized to tabulate the number of potential sources (total exposures). Generalized estimating equations with a logit link were used to identify associations between exposures and progression, and investigate other potential risk factors.

Results The median total exposures was 15 (IQR: 3-23) for cases and 3 (2-12) for controls (p=0.001). The adjusted OR for disease was 1.11 (95% CI 1.06-1.16) per additional exposure, corresponding to an OR of 3.4 for disease when comparing the medians of 15 versus 3 total exposures. This association increased when restricting to TST conversions.

Conclusions Increased exposure may be a marker of greater risk of progression to TB disease. Therefore, this risk may not be transportable across epidemiologic settings with variable exposure intensities.

Introduction

Between November of 2011 and March 2012, there were 23 individuals diagnosed with tuberculosis in a Canadian village of only 933. In response to this crisis, public health and local clinical staff conducted extensive contact investigations of all persons diagnosed with active TB, to identify and treat those with prevalent disease, and identify infected individuals at risk of progression. In all, 50 people were diagnosed with culture-confirmed TB by November 2012 (5% of the village), including 34 of 169 newly infected contacts (20%). This is in stark contrast to the 2-5% risk of progression that has previously been reported for the years immediately following TB infection [1-4]. Our previous studies on housing, nutrition and behavioral characteristics in this community [5, 6], which has a low prevalence of HIV, did not identify factors that could potentially account for this extraordinary rate of disease.

Public health data indicated that newly-infected individuals had been in contact with multiple persons with active TB. Furthermore, molecular epidemiologic analysis revealed that what appeared at the level of the village to be a single 'outbreak' was in fact the result of multiple contemporaneous transmission networks within the same community [7]. Given limited in- and out-migration, and the small size of the community, we could infer that many infected contacts had been repeatedly exposed. These observations led us to hypothesize that the intensity of exposure might explain the elevated risk of active TB disease. To investigate this possibility, we compared newly-infected subjects who progressed to disease with those that did not develop active TB.

Materials and methods

Study design

We conducted a case-control study using a public health database provided by the Nunavik Regional Board of Health and Social Services (NRBHSS). This database includes all individuals with active TB diagnosed in this community between November 2011 – November 2012 and their contacts.

Active TB disease

Individuals with ≥ 1 culture positive for *M. tuberculosis* were defined as 'confirmed TB'. Culture results were assessed for potential cross-contamination as described in [7]. Persons with clinical and radiographic findings consistent with TB, absent culture confirmation, were classified as 'probable' TB.

Contact investigation

Persons with confirmed TB were interviewed upon diagnosis by trained health care providers to obtain lists of household and non-household contacts. Individuals were also asked about attendance at local 'gathering houses', wherein public health suspected transmission might be occurring. These were homes of residents that also served as venues of socialization, as there are no restaurants or bars in this community.

Study inclusion criteria. To assess the proximal risk of progression from infection to active TB, we included only villagers with new TB infection.

New infection

A person was considered to have 'new' TB infection if he/she had a positive TST, either with no previous TST or with a previously documented negative TST (TST conversion).

Cases and Controls

Individuals with new infection who had confirmed TB were included as cases. Contacts with new infection who *did not* progress to active TB in the year following infection were included as controls. A one year follow-up for progression was chosen based on two considerations: 1) the highest proportion of TB disease occurs in the first year following exposure [2, 8] and 2) because of renewed TB transmission in this village in 2014, we sought to avoid confusing 2011-2012 risk factors with those of the next wave of transmission.

Exposure ascertainment

To assess intensity of exposure, we examined the number of times an individual was listed as a contact of a potential source ('total exposures'). Clinical/demographic characteristics of these 50 potential source cases are provided in the online supplementary material (Table S1). Each time a

contact was listed by an individual with active TB, this was counted as one exposure for the contact (i.e., this was done in a unidirectional fashion). This included household and non-household contact. Total exposures also included shared attendance or residence at gathering houses. The precise intensity of contact (i.e., duration and frequency) could not be included in modeling of exposure, as these data were not obtained in a consistent manner throughout the 'outbreak'. Additional details are provided in the Supplementary Material.

One person was never listed as a contact; this person was assigned one total exposure.

Covariates

Data were collected during the 'outbreak' as part of routine contact investigation. Covariates were selected for inclusion from available data based on *a priori* consideration as determinants of TB. These included age at infection, sex, cigarette smoking, residing with a person with smear positive disease, the number of persons per room (as a measure of housing occupancy), Bacillus Calmette-Guerin (BCG) vaccination and HIV/other comorbidities/immunosuppressive disorders. Precise definitions of each and how these were calculated (if applicable) have been provided in the online supplementary material.

Analysis

Main analytic approaches and their respective sample sizes are outlined in Figure 1. In all analyses, the outcome of interest was progression to active TB disease. Our preliminary analysis (analysis 1a) included cases and controls as previously defined. Our secondary analysis (analysis 2a) restricted exposure to contact with individuals with smear positive disease only, in order to assess whether the potential exposure-outcome relationship varied by smear status of potential sources.

We then performed several sensitivity analyses. Firstly, for contacts with a new positive TST that had not been previously tested, it is possible that some of these individuals were already positive before 2011-2012. To address this, we restricted the previous analyses to those with documented TST conversion (analysis 1b and 2b).

Secondly, we repeated the above analyses (1ab, 2ab) using an alternative metric of exposure: the number of different genotypes to which an individual was exposed ('genotypic exposures'). Such exposures were tabulated again based on contact investigation data. The genotypes of each individual with confirmed TB previously identified in [7]. Notably, while there were sufficient genetic differences to assign the multiple chains of transmission, isolates were derived from a strain of *M. tuberculosis* first seen in this village in 2007, therefore it is unlikely that major differences in virulence had developed in the ensuing 6 years [7]. For further details on how these exposures were determined, please see the online supplementary material.

Finally, because the public health response was amplified as of May 1, 2012, with extra clinical staff arriving to assist with contact investigations, we also assessed whether this change influenced results by stratifying analysis 1a by time.

Statistical approaches

Descriptive statistics were conducted and generalized estimating equations with a logit link were used to evaluate the association between potential risk factors and progression to TB disease, accounting for clustering by household. Multiple imputation with chained equations was used to estimate missing data (Table S2). Variables with $p<0.2$ on univariate analysis were assessed in multivariate analysis. Based on previously reported results [6], we evaluated for an interaction between residing with a person with smear positive disease and the number of persons per room; in order to maintain hierarchy, both of these variables were included in preliminary multivariate models *regardless* of significance on univariate analysis. Final models were selected using the Quasi-Information Criterion (QICu) [9]. All analyses were conducted in Stata (v.13, StataCorp 2013).

Ethics

Ethics approval was obtained from the McGill University Faculty of Medicine Institutional Review Board and the NRBHSS. Individual patient consent was not required. All research was done in collaboration with the village council. Databases were linked and analyzed in nominal form, under a professional mandate from the NRBHSS.

Results

Between November 2011 and November 2012, 695 of 933 (74%) villagers, including those with active TB, were investigated by the NRBHSS. Of 169 identified with new infection, 17 individuals were classified as 'probable TB'. These individuals presented distinct characteristics compared to those with confirmed disease (Table S3) and controls (shown in Table 1), and thus they were excluded from analyses. Of the remaining 152 newly infected individuals, 34 had confirmed active TB – 31 with prevalent disease and 3 who developed disease within the year following identification of infection. Two of the latter had agreed to INH prophylaxis, but did not complete therapy. All individuals had pulmonary TB. The remaining 118 subjects were classified as controls.

Summary characteristics of cases with confirmed TB and controls with new infection are shown in Table 1 for analysis 1a. All additional analyses used subsets of these individuals. Three controls were missing address, and were therefore excluded. Overall, cases and controls were similar in terms of age, sex, current cigarette smoking, BCG vaccination status and residing with a person with smear positive disease ($p>0.05$). There was only one person with HIV, who was diagnosed with active TB; otherwise, no relevant comorbidities were identified in either cases or controls. Compared to controls, cases reported higher total exposures ($p=0.001$) and resided in dwellings with higher occupancy, as measured by persons per room ($p=0.036$). Genotypic exposures were also higher for cases compared to controls ($p=0.005$).

Tables 2 and 3 show univariate and multivariate results for analysis 1 and 2, respectively. HIV and other comorbidities were not modeled, due to low/zero cell counts. All continuous variables had linear associations with the outcome ($p>0.05$), except for genotypic exposures in analysis 1a (Table S4). The maximum number of total exposures to any potential source (analysis 1) was 28, while the maximum number of total exposures to sources with smear positive disease only (analysis 2) was 8. Univariate analysis showed a significant association between total exposures and disease, irrespective of type of contact (analysis 1 and 2) or definition of new infection (analysis 1b and 2b). Persons per room was also significantly associated with disease in both analyses.

In all multivariate analyses (1a, 1b, 2a, 2b), total exposures were associated with progression to active TB disease, with adjusted ORs ranging from 1.11-1.53 for each one-person increase in contact (Tables 2 and 3). Persons per room was also significantly associated with progression; the addition of one person to a 5-room dwelling was associated with adjusted ORs from 1.18-1.40 (analysis 1a, 2a, and 2b). When restricted to TST conversions (analysis 1b), an interaction between persons per room and residing with a smear positive was detected; the odds of progression were higher with increased occupancy when a smear positive individual lived in the same residence compared to houses without such individuals. However, as the 95% CIs overlap, this was inconclusive.

Similar results were obtained when the number of different genotypes was used instead as the exposure variable (Tables S4 and S5). There were a maximum of 7 genotypic exposures when considering contact with any potential source versus 6 genotypic when analyses were restricted to smear positive sources only.

To assess the potential influence of clinical staffing changes on May 1, 2012, we conducted separate analyses restricting to contact with persons diagnosed (and therefore interviewed) before or after this date. The number of exposures, and their association with progression to active TB, were similar across time periods (Table 4).

Discussion

Our analysis revealed a significant association between the number of times an individual with recent infection was exposed to active TB and progression to disease. Adjusting for housing occupancy, we found the odds of disease were ~1.1-fold higher for each additional exposure, corresponding to an OR of 3.4 when comparing the median exposures actually experienced by individuals in this community. These results were consistent across all analyses, including when we imposed a more rigorous definition for infection, restricting to TST conversions, and when we restricted our exposure measurement to smear positive sources only. These findings were also unaffected by changes in staffing during the ‘outbreak’.

We propose two possible explanations for the observed association, which need not be mutually exclusive. A first possibility is that the number of exposures is a marker of increased probability of encountering a highly transmissible source. It has been proposed that 20% of all infectious disease cases are responsible for 80% of transmission [10], with the majority of cases either not transmitting at all or very minimally. Such 'super-spreading' has been reviewed in [11] and reported for Severe Acute Respiratory Syndrome, Middle East Respiratory Syndrome and other pathogens. Anecdotal evidence [12, 13] and the high heterogeneity observed in cluster sizes on genotyping [14] suggests this phenomenon also occurs in TB. Our data were consistent with such an explanation, as there were highly contagious cases with smear-positive, cavitary disease within each subgroup of transmission (identified in [7]).

Another possible explanation is that repeated exposures directly influence progression. Increasing exposure (via inoculum size) has been shown to result in more extensive pathology in animals [15-17], however the role dose plays in the development of disease has not been fully elucidated. Experimental human challenges have not been done for TB, nor to our knowledge for other respiratory bacterial pathogens. In other infectious diseases, a dose-response from infection to disease has been reported for *Salmonella* [18, 19], but not *Campylobacter* [20] or *Cryptosporidium* [21, 22]. To ethically investigate this in TB, observational data has been necessary. Previous studies (reviewed in [23]) relying on categorical measures of exposure, such as close versus casual contact [2, 24, 25], household versus non-household [26] or the nature of occupational exposure [27] have also supported an association with progression to disease.

Regardless of the mechanism, we propose that the number of exposures could serve as a useful marker of risk for progression in those with recent infection. Unlike exposure to a super-spreader, which is only known retrospectively following genotyping, the number of times a person is identified as a contact is tabulated in real-time during public health investigations. From a clinical and public health perspective, closer monitoring could be warranted for repeatedly-exposed individuals as they may be at higher risk of progressing to disease.

There were a number of strengths of this study. The dichotomous approach used previously to assess exposure may result in substantial residual confounding. In this study, we were able to obtain continuous metrics of exposure to quantify whether there is an association between increased exposure and risk of disease. Using contact investigation data collected in real-time and building on our molecular epidemiologic analyses [7], we tabulated the total number of exposures to different potential sources of TB, as well as the number of genotypic exposures. For both metrics, there were increased odds of progression from infection to disease with higher exposure. An additional strength was the limited out-migration from this region, which facilitated collection of complete contact investigation data and 100% follow-up for progression to active TB. Finally, as most persons had resided in the village since birth, we also had access to complete, life-long medical records to assess for comorbidities and TB risk factors.

This study has several limitations. Sample size was limited by the extent of the public health crisis, with 34 individuals diagnosed with active TB among those with new infection during the study period. This may have reduced our power to detect associations between other covariates and progression to disease. Data were also collected as part of the public health response to the outbreak, rather than to test specific research hypotheses. As such, we could only assess variables routinely collected during a TB control investigation. Standardized contact investigation tools ensured that key covariates such as age and sex were consistently documented, however, and we note that follow-up studies in this village have similarly reported lack of association between these covariates and progression to active TB [5, 6]. Unlike most environments where contact investigation has been studied, the epidemiologic context is more homogeneous; as >90% of villagers were Inuit [28], residing in the same isolated Northern community, many social characteristics were similar, potentially reducing the ability detect associations with disease. Smoking, for example, is quite prevalent in this community; without an exposure gradient, we could not detect an association in this context, despite smoking being linked to TB disease in many other populations. In accordance with the previous studies in this village, we found that housing occupancy was associated with progression, thereby strengthening confidence in our results. Finally, while the small size of the community and limited migration has made it feasible to detect all cases and subsequently perform large-scale contact

investigations, this may not be as feasible in other settings with greater migration and loss to follow-up.

Through this analysis, we have shown that multiple exposures to TB are associated not only with increased infection, but increased progression to disease as well. From a public health standpoint, such exposures could therefore serve as a marker of increased risk of progression to disease. Given the unique nature of this outbreak, these findings need to be validated in other settings. If exposure intensity is a marker of progression to TB disease, then attack rates from low-prevalence settings may under-estimate the risk of disease in settings where multiple exposures are more likely, and vice-versa.

Acknowledgements

Authors thank the village council and residents for their collaboration and engagement in this study. Authors also thank the staff of the Local Community Service Centre for their hard work during the outbreak, including collection of contact investigation data, and the Nunavik Regional Board of Health and Social Services for provision of clinical and epidemiologic data used in this study. Authors also thank David C. Alexander, Director of Virology, Saskatchewan Disease Control Laboratory, Dr. Brian Ward, Associate Professor at the Centre for the Study of Host Resistance, McGill University Health Centre and Dr. Cedric Yansouni, Associate Professor, Faculty of Medicine, McGill University, for sharing historical papers.

Financial support

This work was supported by the Canadian Institutes of Health Research [MOP number 125858, to DM and MAB]. This funding agency had no role in the study design, data collection, analysis and interpretation of data, or in the writing the manuscript or the decision to submit for publication.

FIGURES

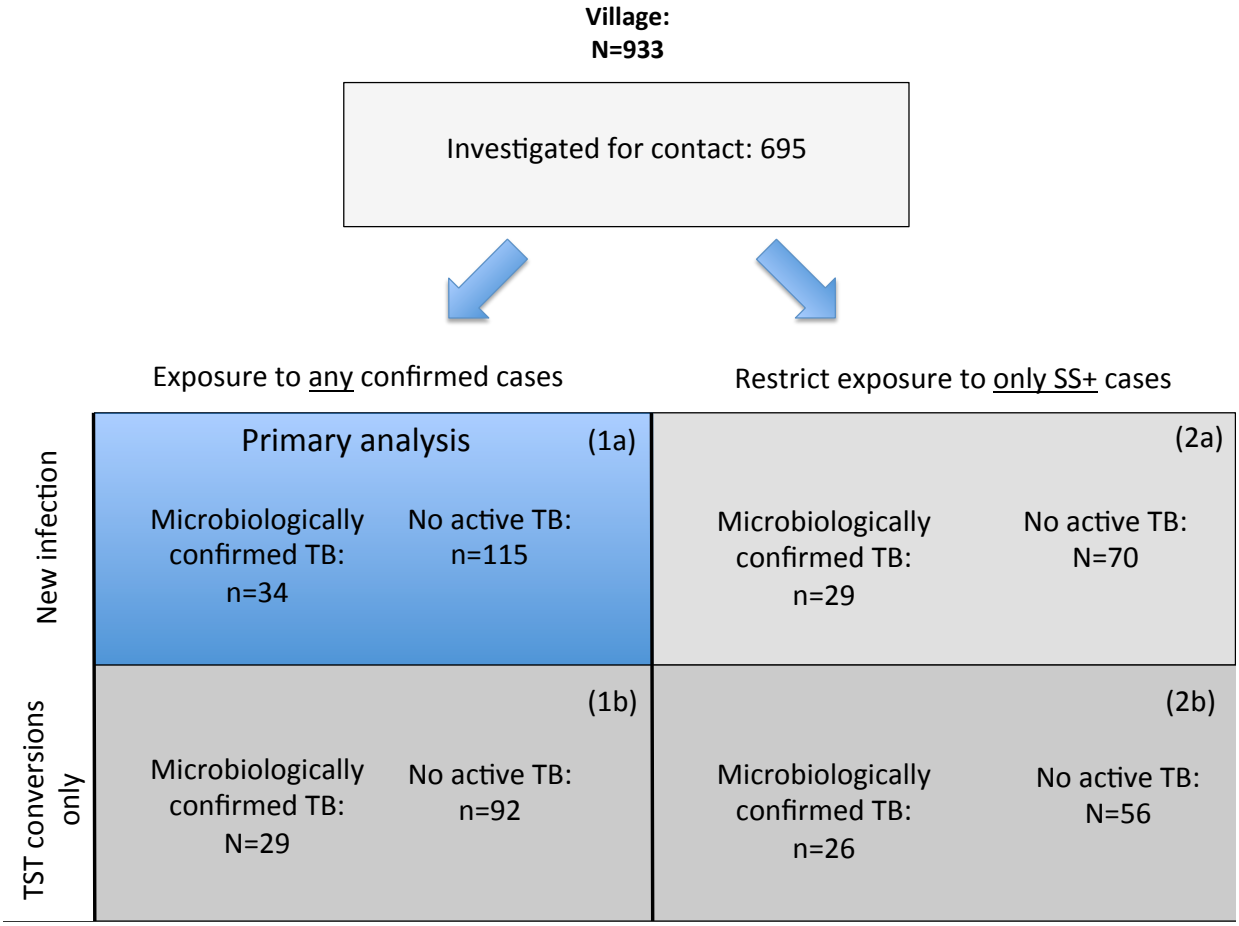


FIGURE 7-1. Main analytic approaches.
The number of cases and controls included in each analysis are indicated. Note that 3 controls were excluded from analyses due to missing address. Fewer subjects are also present in analysis 2 as some individuals did not have contact with a smear positive source. One person who did not have any reported contact was assigned a single total (and single genotypic) exposure as a minimum; this individual was not included in the smear positive analysis as the contagiousity of his/her potential source was unknown.

TABLES

TABLE 7-1. Characteristics of individuals with exposure to any potential source, new infection (analysis 1a).

Variables of interest	Confirmed TB disease (cases, n=34)	No disease (controls, n=115)*	p value
Age at infection, median (interquartile range, IQR), y	19.5 (15.3-28.1)[$]	19.9 (12.4-25.5)	0.654[†]
No. (%) under 5 y age	*4 (12)*	*6 (5)*	*0.237*[#]
No. (%) male sex	18 (53)	63 (55)	0.850[‡]
No. (%) current smoking	23 (77)	65 (65)	0.246[‡]
No. (%) Bacillus Calmette-Guerin (BCG)	25 (76)	96 (83)	0.192[‡]
No. (%) residing with a person with smear positive disease	7 (21)	13 (11)	0.163[‡]
No. (%) co-morbidities (HIV, diabetes, renal dysfunction, other immunosuppressive disorders)	1 (3)	0 (0)	0.228[#]
Total exposures, median (IQR)	15 (3-23)	3 (2-12)	0.001[†]
Genotypic exposures, median (IQR)	5 (2-6)	2 (1-4)	0.005[†]
Persons per room, median (IQR)	1.8 (1.3-2.7)	1.7 (1.2-2.3)	0.036[†]

*Three controls excluded from analysis as missing address of residence. [$]Age range for cases: 1.1-54.8 years, age range for controls: 0.5-59.4 years). [†]Mann-Whitney ranksum test. [‡]Chi-square test with 2 degrees of freedom. [#]Fisher's Exact test. Non-missing data are used for the denominator of proportions. A two-sided p value of <0.05 is considered statistically significant.

305

TABLE 7-2. Exposure to any potential source and progression to active TB.

306

		Univariate		Multivariate	
	Odds ratio	**95% CI**	**p value**	**Odds ratio**	**95% CI**
Analysis 1a – Contact with any potential source, newly diagnosed infection					
Age at infection	1.00	0.97-1.03	0.806	Not in final model	
Male sex	0.93	0.44-1.94	0.844	Not in final model	
Current smoking	1.63	0.58-4.54	0.351	Not in final model	
BCG	0.65	0.24-1.75	0.391	Not in final model	
Residing with a person with smear positive disease	2.03	0.67-6.14	0.208	Not in final model	
Total exposures	1.09	1.05-1.14	<0.0005	1.11	1.06-1.16
Persons per room*	1.12	0.98-1.28	0.086	1.18	1.04-1.34
Analysis 1b - Contact with any potential source, tuberculin skin test conversion only					
Age at infection	1.01	0.98-1.04	0.593	Not in final model	
Male sex	0.72	0.31-1.67	0.442	Not in final model	
Current smoking	2.19	0.61-7.90	0.231	Not in final model	
BCG	1.30	0.33-5.21	0.707	Not in final model	
Residing with a person with smear positive disease	3.17	0.85-11.79	0.085	0.27	0.02-4.59
Total exposures	1.12	1.06-1.18	<0.0005	1.14	1.08-1.21
Persons per room*	1.13	1.00-1.28	0.056		
Persons per room* *when **not** residing with a person with smear positive disease*				1.15	1.03-1.28
Persons per room* *when residing with a person with smear positive disease*				1.49[†]	1.06-2.10

307 *For comparability to [6]. Persons per room scaled such that odds ratio corresponds to a 1 person increase
308 in a 5-person house. [†]p=0.027 for interaction between persons per room and residing with a person with
309 smear positive disease; this OR represents the joint effect of adding 1 person to a 5-person house when
310 residing with an individual with smear positive disease. Age and persons per room are centered at the
311 overall mean for analysis 1a, at 20.8 years and 1.7 persons per room, respectively. Total exposures centered
312 at 1, as all individuals had at least 1 contact.
313

TABLE 7-3. Exposure to potential sources with smear positive disease only and progression to active TB.

		Univariate		Multivariate	
	Odds ratio	**95% CI**	**p value**	**Odds ratio**	**95% CI**
Analysis 2a – Potential sources with smear positive disease only, newly diagnosed infection					
Age at infection	1.01	0.97-1.05	0.603	Not in final model	
Male sex	1.23	0.54-2.81	0.623	Not in final model	
Current smoking	1.26	0.36-4.38	0.719	Not in final model	
BCG	0.93	0.23-3.71	0.922	Not in final model	
Residing with a person with smear positive disease	1.40	0.44-4.43	0.572	Not in final model	
Total exposures	1.34	1.11-1.60	0.002	1.44	1.18-1.76
Persons per room*	1.24	1.09-1.41	0.001	1.33	1.15-1.54
Analysis 2b – Potential sources with smear positive disease only, tuberculin skin test conversion only					
Age at infection	1.01	0.97-1.06	0.574	Not in final model	
Male sex	1.00	0.40-2.52	1.000	Not in final model	
Current smoking	1.42	0.36-5.64	0.621	Not in final model	
BCG	1.28	0.23-6.95	0.778	Not in final model	
Residing with a person with smear positive disease	2.10	0.54-8.15	0.284	Not in final model	
Total exposures	1.39	1.14-1.71	0.001	1.53	1.24-1.88
Persons per room*	1.29	1.12-1.48	0.001	1.40	1.18-1.66

For comparability to [6]. Persons per room scaled such that odds ratio corresponds to a 1 person increase in a 5-person house. Age and persons per room are centered at the overall mean for analysis 1a, at 20.8 years and 1.7 persons per room, respectively. Total exposures centered at 1, as all individuals had at least 1 contact.

TABLE 7-4. Exposure to any potential source and progression to active TB (analysis 1a), stratified by time of diagnosis of the source

	Cases			Controls	
November 2011-April 2012					
N	31			85	
Total exposures, median (IQR)	8 (2-10)			2 (1-7)	
May 2012-November 2012					
N	29			98	
Total exposures, median (IQR)	9 (2-15)			3 (1-6)	
		Univariate		**Multivariate**	
	Odds ratio	**95% CI**	**p value**	**Odds ratio**	**95% CI**
November 2011-April 2012					
Age at infection	1.01	0.97-1.04	0.759	Not in final model	
Male sex	0.75	0.34-1.64	0.466	Not in final model	
Current smoking	1.36	0.46-3.98	0.577	Not in final model	
BCG	0.76	0.25-2.31	0.627	Not in final model	
Residing with a person with smear positive disease	1.96	0.63-6.16	0.248	Not in final model	
Total exposures	1.08	1.04-1.14	0.001	1.10	1.05-1.15
Persons per room*	1.10	0.96-1.26	0.186	1.16	1.01-1.32
May 2012-November 2012					
Age at infection	1.01	0.98-1.05	0.352	Not in final model	
Male sex	1.03	0.46-2.28	0.945	Not in final model	
Current smoking	2.12	0.67-6.72	0.202	Not in final model	
BCG	0.75	0.23-2.50	0.642	Not in final model	
Residing with a person with smear positive disease	2.06	0.61-7.03	0.208	Not in final model	
Total exposures	1.11	1.06-1.16	<0.0005	1.12	1.07-1.18
Persons per room*	1.12	0.98-1.28	0.086	1.18	1.04-1.33

For comparability to [6]. Persons per room scaled such that odds ratio corresponds to a 1 person increase in a 5-person house. Age and persons per room are centered at the overall mean for analysis 1a, at 20.8 years and 1.7 persons per room, respectively. Total exposures centered at 1, as all individuals had at least 1 contact.

328 **References**

329

330 1. Mack U, Migliori GB, Sester M, Rieder HL, Ehlers S, Goletti D, Bossink A,
331 Magdorf K, Holscher C, Kampmann B, Arend SM, Detjen A, Bothamley G,
332 Zellweger JP, Milburn H, Diel R, Ravn P, Cobelens F, Cardona PJ, Kan B,
333 Solovic I, Duarte R, Cirillo DM, C. Lange for the TBNET. LTBI: latent
334 tuberculosis infection or lasting immune responses to M. tuberculosis? A
335 TBNET consensus statement. Eur Respir J 2009; 33: 956–973.

336 2. Sloot R, Schim van der Loeff MF, Kouw PM, Borgdorff MW. Risk of
337 tuberculosis after recent exposure. A 10-Year follow-up study of contacts in
338 Amsterdam. Amer J Respir Crit Care Med 2014; 190: 1044–1052.

339 3. Downes J. A study of the risk of attack among contacts in tuberculous
340 families in a rural area. Am J Epi 1935.

341 4. Public Health Agency of Canada. Canadian Tuberculosis Standards 7th
342 Edition. 2014.

343 5. Fox GJ, Lee RS, Lucas M, Ahmad Khan F, Proulx J-F, Hornby K, Jung S,
344 Benedetti A, Behr MA, Menzies D. Inadequate diet is associated with
345 acquiring *Mycobacterium tuberculosis* infection in an Inuit community: A
346 case-control study. Annals ATS 2015; 12(8): 1153-1162.

347 6. Ahmed Khan F, Fox GJ, Lee RS, Riva M, Benedetti A, Proulx JF, Jung S,
348 Hornby K, Behr MA, Menzies D. Housing characteristics as determinants of
349 tuberculosis in an Inuit community: a case-control study. 19th Annual
350 Conference of The International Union Against Tuberculosis and Lung
351 Disease - North America Region Vancouver; 2015.

352 7. Lee RS, Radomski N, Proulx J-F, Manry J, McIntosh F, Desjardins F,
353 Soualhine H, Domenech P, Reed MB, Menzies D, Behr MA. Reemergence
354 and amplification of tuberculosis in the Canadian arctic. J Infect Dis 2015;
355 211: 1905–1914.

8. Fox GJ, Barry SE, Britton WJ, Marks GB. Contact investigation for tuberculosis: a systematic review and meta-analysis. Eur Respir J 2012; 41: 140–156.

9. Pan W. Akaike's information criterion in generalized estimating equations. Biometrics 2001; 57: 120–125.

10. Woolhouse ME, Dye C, Etard JF, Smith T, Charlwood JD, Garnett GP, Hagan P, Hii JL, Ndhlovu PD, Quinnell RJ, Watts CH, Chandiwana SK, Anderson RM. Heterogeneities in the transmission of infectious agents: implications for the design of control programs. Proc Natl Acad Sci USA. 1997; 94: 338–342.

11. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. Nature 2005; 438: 355–359.

12. Kiers A, Drost AP, van Soolingen D, Veen J. Use of DNA fingerprinting in international source case finding during a large outbreak of tuberculosis in The Netherlands. Int J Tuberc Lung Dis 1997; 1: 239–245.

13. Kline SE, Hedemark LL, Davies SF. Outbreak of tuberculosis among regular patrons of a neighborhood bar. New Engl J Med 1995; 333: 222–227.

14. Ypma RJF, Altes HK, van Soolingen D, Wallinga J, van Ballegooijen WM. A Sign of Superspreading in Tuberculosis. Epidemiology 2013; 24: 395–400.

15. Koch R, codell Carter K. Essays of Robert Koch. Greenwood Press; 1987.

16. Ratcliffe HL. Tuberculosis induced by droplet nuclei infection. Am J Epi; 1952. p. 30–48.

17. Perla D. Experimental Epidemiology of tuberculosis. J. Exp. Med. 1927; 45: 209–226.

382 18. McCullough N, Eisele W. Experimental human salmonellosis. III.
383 Pathogenicity of strains of *Salmonella newport*, *Salmonella derby* and
384 *Selmonella bareilly* obtained from spray-dried whole egg. J Infect Dis 1951;
385 88: 278–289.

386 19. Blaser MJ, Newman LS. A Review of human salmonellosis: I. Infective
387 dose. Rev. Infect. Dis. 1982; 4: 1096–1105.

388 20. Black RE, Levine MM, Clements ML, Hughes TP, Blaser MJ. Experimental
389 *Campylobacter jejuni* infection in humans. J Infect Dis 1988; 157: 472–479.

390 21. Chappell CL, Okhuysen PC, Langer-Curry RC, Lupo PJ, Widmer G, Tzipori
391 S. *Cryptosporidium muris*: infectivity and illness in healthy adult volunteers.
392 *Am J Trop Med Hyg* 2015; 92: 50–55.

393 22. DuPont HL, Chappell CL, Sterling CR, Okhuysen PC, Rose JB, Jakubowski
394 W. The infectivity of *Cryptosporidium parvum* in healthy volunteers. New
395 Engl J Med 1995; 332: 855–859.

396 23. Salgame P, Geadas C, Collins L, Jones-López E, Ellner JJ. Latent
397 tuberculosis infection--Revisiting and revising concepts. Tuberculosis 2015;
398 95: 373–384.

399 24. Grzybowski S, Barnett GD, Styblo K. Contacts of cases of active pulmonary
400 tuberculosis. Bull Int Union Tuberc Lung Dis 1975; 50: 90–106.

401 25. Houk VN, Baker JH, Sorensen K, Kent DC. The epidemiology of
402 tuberculosis infection in a closed environment. Arch Environ Health: Int J
403 1968; 16: 26–35.

404 26. Moran-Mendoza O, Marion SA, Elwood K, Patrick D, FitzGerald JM. Risk
405 factors for developing tuberculosis: a 12-year follow-up of contacts of
406 tuberculosis cases. Int J Tuberc Lung Dis 2010; 14: 1112–1119.

407 27. Ferguson RG. BCG vaccination in hospitals and sanatoria of Saskatchewan;

a study carried out by the National Research Council of Canada. Can J Public Health 1946; 37: 435–451.

28. Statistics Canada. Kangiqsualujjuaq, Quebec (Code 2499090). National Household survey. Statistics Canada catalogue no. 99-011-X20110007. http://www12.statcan.gc.ca/nhs-enm/2011/dp-pd/aprof/index.cfm?Lang=E. Date last updated: November 27 2015. Date last accessed: May 4 2016.

CHAPTER 8. OBJECTIVE 4 – Manuscript IV

Lee RS and Behr MA. Does choice matter? Reference-based alignment for molecular epidemiology of tuberculosis. Accepted *J Clin Micro* on April 5, 2016.

8.1 Preamble

In Manuscripts I and II, WGS was utilized to investigate TB transmission. As this method has only recently become feasible – for both cost and technical reasons – there is currently no standardized approach to data analysis. As epidemiologists rely on these SNPs to discern relationships between isolate, it is critical to understand how different analytic decisions made can influence these results. The following manuscript therefore examines one of the bioinformatics decisions for analysis of *M. tuberculosis*.

Additional analyses in the form of supplementary data (accepted for publication) can be found in **Appendix 5**.

8.2 Manuscript IV

Does choice matter? Reference-based alignment for molecular epidemiology of tuberculosis

Robyn S. Lee[a-c] and Marcel A. Behr[b-d#]

McGill University, Department of Epidemiology, Biostatistics and Occupational Health[a]; The Research Institute of the McGill University Health Centre[b]; McGill International TB Centre[c]; McGill University Health Centre, Department of Medicine, Division of Infectious Diseases[d]

Running Head: Choosing a reference for *Mycobacterium tuberculosis*

#Address correspondence to Marcel A. Behr, marcel.behr@mcgill.ca.

Abstract

When using genome sequencing for molecular epidemiology, short sequence reads are aligned to an arbitrary reference strain to detect single nucleotide polymorphisms. We investigated whether reference genome selection influences epidemiologic inferences of *Mycobacterium tuberculosis* transmission, by aligning sequence reads from 162 closely-related Lineage 4 (Euro-American) isolates to 7 different genomes. Phylogenetic trees were consistent using all but the most divergent genomes, suggesting that reference choice can be based on considerations other than *M. tuberculosis* lineage.

Whole genome sequencing (WGS) has become the gold standard for molecular epidemiology studies of *Mycobacterium tuberculosis*, demonstrating higher resolution than classical molecular typing methods (e.g.,(1-5)). Epidemiologic inferences depend on the detection of single nucleotide polymorphisms (SNPs) that distinguish isolates. Identifying SNPs using short-read data typically involves alignment ('mapping') of reads to a single reference genome (e.g., *M. tuberculosis* H37Rv). As the difference between the genome of the reference strain and the clinical isolates increases (e.g. insertions / deletions / SNPs), fewer sequence reads are successfully mapped against the reference genome. As these data are essentially lost, the results are potentially biased and true differences may go undetected. One solution in studies of other bacterial pathogens has been *de novo* assembly of a closely-related isolate; this is then used in lieu of existing, more genetically distant reference genomes (6). However, this approach requires additional resources, in terms of cost, technical expertise and time; if short-read data is used for *de novo* assembly, a much greater sequencing depth is required (>100x (7)) to ensure sufficient overlap of reads to facilitate accurate assembly, while alternative sequencing platforms are necessary to generate longer reads.

We asked whether the use of different reference genomes influences phylogenetic trees and epidemiologic inferences of *M. tuberculosis* transmission, utilizing an existing dataset of 163 Lineage 4 (Euro-American) isolates from Northern Quebec. DNA extraction and MiSeq-based WGS were performed as previously described ((8), National Center for Biotechnology Information's Sequence Read Archive Accession SRP039605, Bioproject PRJNA240330). Mixed infection with *Mycobacterium avium* was identified in

1 isolate using the Basic Local Alignment Search Tool (9); while this had no influence on previous phylogenies, it was excluded from the current analysis to avoid bias in coverage calculations (below). Read quality was assessed with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were trimmed using Trimmomatic (v.0.32 (10)), with minimum length of 70 base-pairs (bp), then aligned using the Burrows-Wheeler Aligner (BWA) MEM algorithm (11) to 7 different reference genomes (**Table 8-1**, divergence in average nucleotide identity given in **Table S1**). PCR and optical duplicates were marked using PicardTools (v.1.118, available at http://broadinstitute.github.io/picard/) and reads were locally re-aligned around insertions/deletions (indels). Reads aligning to >1 locus in the reference, or with mapping quality <30 were excluded. The proportion of reads that aligned to each reference was calculated using Samtools (v.1.2, (12)). Genome coverage and average depth of coverage were calculated excluding duplicates in QualiMap (v.2, (13) and Integrative Genomics Viewer (14) (**Tables S2** and **S3**).

The highest proportion of reads mapped to the CDC1551 (Lineage 4) reference, followed by H37Rv (**Table 8-1**). As both are lineage 4, this is unsurprising. The mean proportions of the CDC1551 and H37Rv references that had at least 1 read aligned ('genome coverage') were also highest across all analyses. As the reference strain became more genetically divergent from the sequenced isolates (Lineage 2 *M. tuberculosis*, *Mycobacterium africanum*, *Mycobacterium bovis* and *Mycobacterium canettii*), a minor decline in the percent of total reads aligned and genome coverage was evident. When

aligning against *Mycobacterium kansasii*, these values decreased by 45.5% and 64.8%, respectively, when compared to CDC1551.

SNPs and indels were then identified ('called') for each reference analysis using the Genome Analysis Toolkit (GATK v.3.3, (15)). SNPs were filtered for quality based on GATK recommendations, including assessment of strand bias. In addition, we required a Phred ≥50 (where Phred = -10*log P_{error}, corresponding to a 1/100,000 probability of error) for all loci, a minimum depth of coverage (i.e., the number of reads that are aligned to that locus) of 8 bp and individual Phred-scaled genotype quality ≥15 to confidently call a SNP. SNPs within 12 bp of one another or indels and heterozygous calls were excluded. Concatenated SNPs from each alignment were then used to generate phylogenetic trees using the maximum likelihood method (16) with 1000 bootstrap replicates (17). The model of nucleotide substitution was chosen based on the Bayesian Information Criterion. As repetitive PE_PGRS, PPE genes and mobile elements were not consistently annotated across all reference genomes, SNPs in these regions were included; however any bias due to these SNPs should be non-differential across references. Trees from each analysis were compared qualitatively and were largely consistent with a previous, deletion-based phylogeny (8). As illustrated in **Fig. 8-1** and **S1**, small changes in clustering became evident at the level of *M. canettii*, while resolution was almost entirely lost with *M. kansasii.*

To examine whether reference choices influenced our interpretation of direct patient-to-patient transmission, we restricted our analysis to 49 isolates from a well-defined

epidemiologic 'outbreak' in a single Quebec community. All cases were diagnosed within a 1-year time period and previous work (5) suggested a threshold for recent, direct transmission of 0-1 SNP. Matrices of pairwise SNPs between isolates were generated. Using classifications with CDC1551 as the gold standard, due to its closest genetic similarity to our isolates, we calculated the sensitivity and specificity for classifying each pair as 'probable recent transmission', or not. As shown in **Table 8-2**, the sensitivity and specificity for detecting recent transmission was 100% across all reference genomes, excepting *M. kansasii.* In the latter, nearly all SNPs that formerly ruled out transmission between some pairs were missed because of low mapping to the reference, yielding an unacceptably high number of false positives.

Overall, we have shown that that the choice of reference genome – within the *M. tuberculosis* complex – has negligible influence on phylogeny and epidemiologic studies of *M. tuberculosis* transmission. Because we were able to demonstrate the robustness of these analyses using a dataset with very limited strain diversity (153/163 isolates were separated by a maximum distance of 72 SNPs and clusters were distinguished by as few as 2 SNPs (5, 8), this suggests our findings are generalizable to settings with greater genetic diversity and robust to differences in *M. tuberculosis* lineage. Therefore, epidemiologic studies of TB can base reference choices on aspects such as quality of annotation, rather than matching strain lineage.

Our findings also indicate that there is a threshold of genome coverage beyond which transmission can no longer be accurately discriminated. This can particularly have

implications for non-clonal pathogens, which have greater genetic diversity than *M. tuberculosis*. One approach with such organisms restricts short read alignment to the core genome region (e.g. in *Escherichia coli*, this represents only 40% of all possible genes (18)), while another restricts to variation within pre-selected genes (e.g. 'housekeeping genes' used for multilocus sequence typing). These subsets are then used to build phylogenetic trees and delineate clusters of transmission. When limiting to only a subset of the genome, epidemiologically-relevant genetic diversity can be overlooked, as demonstrated when aligning to *M. kansasii*. A more optimal approach might involve aligning to both core and accessory genes and >1 reference from the same species, to capture a more complete portrait of bacterial diversity. To facilitate this, efforts must be made to further sequence, close and annotate such genomes.

FIGURES

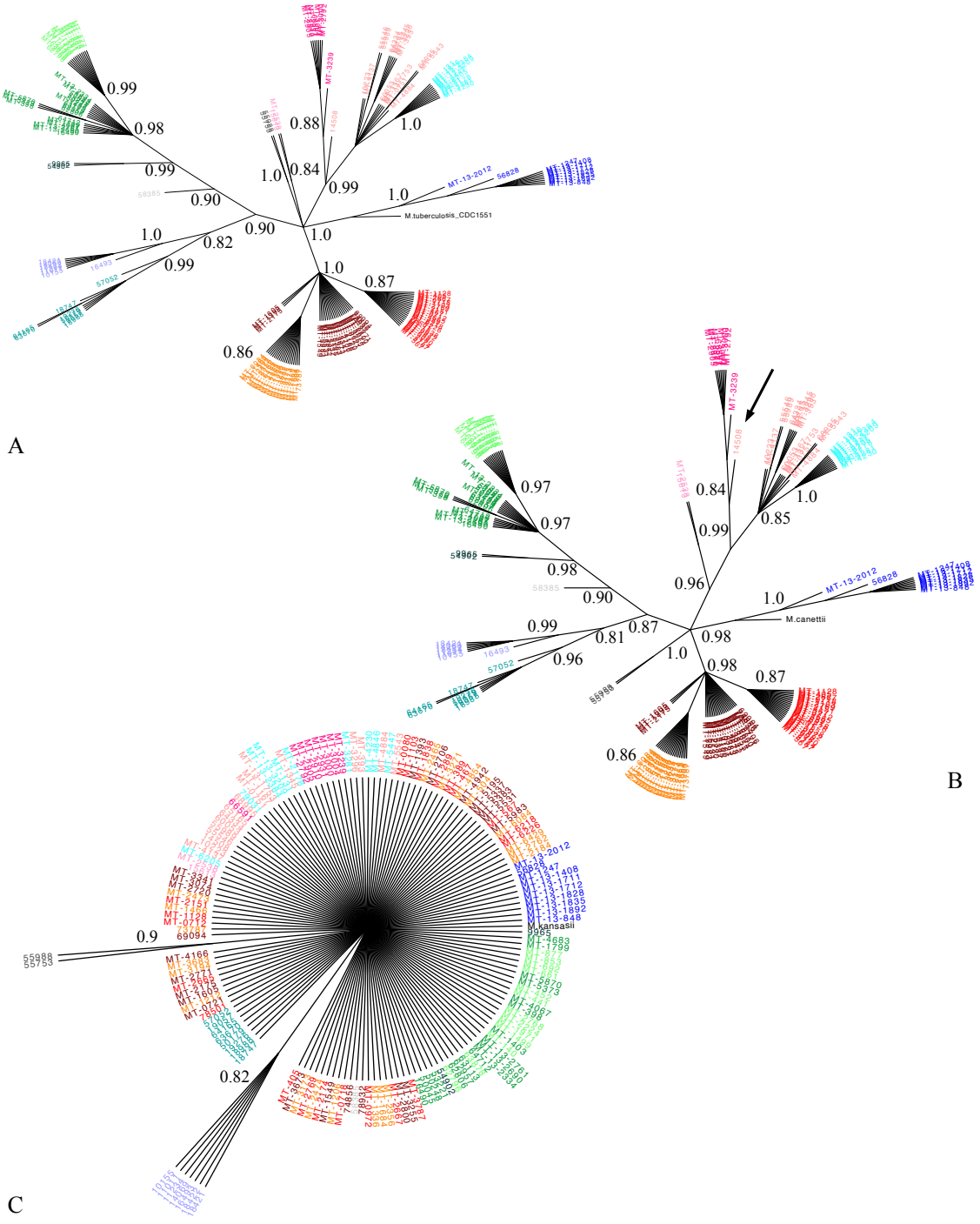


FIGURE 8-1. Impact of reference genome choice on phylogeny.
Maximum likelihood trees with 1000 bootstrap replicates. Branches below 80% bootstrap threshold are collapsed (branch lengths are therefore not to scale). For clarity, bootstrap p values are indicated up until the most proximal node defining each cluster. Isolates were

coloured for their respective clusters identified according to CDC1551 (and H37Rv (8)). Isolates were then kept the same colour across all panels, to facilitate quick comparison between the new reference analysis and CDC1551. See **Table S4** for cluster names. A – Reference *M. tuberculosis* Lineage 4 CDC1551, using the Tamura 3-parameter (19) model of nucleotide substitution with 1,522 SNP loci. B – Reference *M. canettii*, using the GTR model of nucleotide substitution (20) with 17,406 SNP loci. Using *M. canettii* as a reference, a single isolate changed clusters, indicated with an arrow. C – Reference *M. kansasii*, using the GTR model of nucleotide substitution with 34,127 SNP loci.

TABLES

TABLE 8-1 Alignment and genome coverage across various reference genomes within the genus *Mycobacteria.*

Reference genome species	Reference genome name	Accession number	Citation	Reference genome length (base pairs)	Percent of reads successfully [a] aligned to reference, median (IQR)	Genome coverage at ≥ 1x depth, median (IQR) [b]	Genome coverage at ≥ 10x depth, median (IQR) [b]	Genome coverage at ≥ 20x depth, median (IQR) [b]
Mycobacterium tuberculosis, lineage 4	H37Rv	NC_000962.3	(21)	4,411,532	98.0 (97.9-98.1)	98.9 (98.8-98.9)	98.1 (98.0-98.3)	97.2 (96.8-97.5)
Mycobacterium tuberculosis, lineage 4	CDC1551	NC_002755.2	(22)	4,403,837	98.2 (98.1-98.3)	99.3 (99.2-99.3)	98.5 (98.4-98.7)	97.6 (97.2-97.9)
Mycobacterium tuberculosis, lineage 2	CCDC5079	CP001641	(23)	4,398,812	97.8 (97.7-97.9)	98.8 (98.7-98.8)	98.1 (97.9-98.2)	97.1 (96.8-97.4)
Mycobacterium africanum	GN041182	FR878060.1	(24)	4,389,314	97.5 (97.4-97.5)	98.9 (98.8-98.9)	98.1 (98.0-98.3)	97.2 (96.8-97.4)
Mycobacterium bovis	AF2122/97	NC_002945.3	(25)	4,345,492	97.6 (97.5-97.7)	99.3 (99.2-99.3)	98.5 (98.4-98.7)	97.6 (97.2-97.9)
Mycobacterium canettii	CIPT 140010059	NC_015848.1	(26)	4,482,059	96.5 (96.3-96.6)	95.1 (95.0-95.1)	94.3 (94.2-94.4)	93.4 (93.0-93.8)
Mycobacterium kansasii [d]	ATCC 12478	NC_022663.1	(27)	6,432,277	52.7 (51.6-53.4)	34.5 (34.2-35.5)	28.4 (27.8-29.1)	25.5 (24.6-26.4)

[a] Calculated using Samtools –flagstat- as (total mapped – secondary alignments – duplicate reads)/(total reads surviving trimming - duplicate reads).
[b] QualiMap includes secondary alignments marked by BWA MEM (range: 1-3% of total mapped), double-counted in coverage calculations. Duplicates excluded.
[c] pMK plasmid sequence not used for alignment.

TABLE 8-2 Comparing pairwise single nucleotide polymorphisms (SNPs) and probable recent transmission by reference genome, using CDC1551 as the gold standard.

Reference genome species	Reference genome name	Median pairwise SNPs compared to the reference (IQR) [a]	Median pairwise SNPs between isolates (IQR) [b]	Sensitivity for recent transmission (95% CI)	Specificity for recent transmission (95% CI)
Mycobacterium tuberculosis, lineage 4	H37Rv	781 (780-781)	3 (2-6)	100 (98.7-100)	100 (99.6-100)
Mycobacterium tuberculosis, lineage 4	CDC1551	619 (618-619)	3 (2-6)	-	-
Mycobacterium tuberculosis, lineage 2	CCDC5079	1,247 (1,246-1,247)	3 (2-6)	100 (98.7-100)	100 (99.6-100)
Mycobacterium africanum	GN041182	1,908 (1,907-1,908)	3 (2-6)	100 (98.7-100)	100 (99.6-100)
Mycobacterium bovis	AF2122/97	2,000 (1,999-2,000)	3 (2-6)	100 (98.7-100)	100 (99.6-100)
Mycobacterium canettii	CIPT 140010059	16,637 (16,636-16,637)	3 (2-6)	100 (98.7-100)	100 (99.6-100)
Mycobacterium kansasii	ATCC 12478	34,081 (34,081-34,081)	0 (0-0)	100 (98.7-100)	0.1 (0.0-0.06)

[a] 49 pairwise comparisons with the reference genome. [b] 1,176 pairwise comparisons.

References

1. **Niemann S, Köser CU, Gagneux S, Plinke C, Homolka S, Bignell H, Carter RJ, Cheetham RK, Cox A, Gormley NA, Kokko-Gonzales P, Murray LJ, Rigatti R, Smith VP, Arends FPM, Cox HS, Smith G, Archer JAC.** 2009. Genomic diversity among drug sensitive and multidrug resistant isolates of *Mycobacterium tuberculosis* with identical DNA fingerprints. PLoS One **4**:e7407. doi:10.1371/journal.pone.0007407

2. **Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L, Brodkin E, Rempel S, Moore R, Zhao Y, Holt R, Varhol R, Birol I, Lem M, Sharma MK, Elwood K, Jones SJM, Brinkman FSL, Brunham RC, Tang P.** 2011. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N Engl J Med **364**:730-739.

3. **Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, Eyre DW, Wilson DJ, Hawkey PM, Crook DW, Parkhill J, Harris D, Walker AS, Bowden R, Monk P, Smith EG, Peto TE.** 2013. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. Lancet Infect Dis **13**:137-146.

4. **Lee RS, Radomski N, Proulx JF, Manry J, McIntosh F, Desjardins F, Soualhine H, Domenech P, Reed MB, Menzies D, Behr MA.** 2015. Reemergence and amplification of tuberculosis in the Canadian arctic. J Infect Dis **211**(12):1905-14.

5. **Roetzer A, Diel R, Kohl TA, Rückert C, Nübel U, Blom J, Wirth T, Jaenicke S, Schuback S, Rüsch-Gerdes S, Supply P, Kalinowski J, Niemann S.** 2013. Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiological study. PLoS Med **10**:e1001387. doi:10.1371/journal.pmed.1001387

6. **Bryant JM, Grogono DM, Greaves D, Foweraker J, Roddick I, Inns T, Reacher M, Haworth CS, Curran MD, Harris SR, Peacock SJ, Parkhill J, Floto RA.** 2013. Whole-genome sequencing to identify transmission of *Mycobacterium abscessus* between patients with cystic fibrosis: a retrospective cohort study. Lancet **381**:1551-1560.

7. **Ekblom R, Wolf JBW**. 2014. A field guide to whole-genome sequencing, assembly and annotation. Evol Appl **7**(9):1026-1042.

8. **Lee RS, Radomski N, Proulx J-F, Levade I, Shapiro BJ, McIntosh F, Soualhine H, Menzies D, Behr MA.** 2015. Population genomics of *Mycobacterium tuberculosis* in the Inuit. Proc Natl Acad Sci U S A **112**(44):13609-13614.

9. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.** 1990. Basic local alignment search tool. J Mol Biol **215**:403-410.

10. **Bolger AM, Lohse M, Usadel B.** 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30**:2114-2120.

11. **Li H.** 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv http://arxiv.org/abs/1303.3997

12. **Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup.** 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**:2078-2079.

13. **Okonechnikov K, Conesa A, García-Alcalde F.** 2015. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics **32**(2):292-4.

14. **Thorvaldsdottir H, Robinson JT, Mesirov JP.** 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform **14**(2):178–192.

15. **McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA.** 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res **20**:1297-1303.

16. **Felsenstein J**. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol **17**:368-376.

17. **Felsenstein J.** 1985. Confidence-limits on phylogenies - an approach using the bootstrap. Evolution **39**:783-791.

18. **Mira A, Martín-Cuadrado AB, D'Auria G, Rodríguez-Valera F.** 2010. The bacterial pan-genome: a new paradigm in microbiology. Int Microbiol **13**:45-57.

19. **Tamura K.** 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. Mol Biol Evol **9**:678-687.

20. **Waddell PJ, Steel MA.** 1997. General time-reversible distances with unequal rates across sites: mixing Γ and inverse gaussian distributions with invariant sites. Mol Phylogenet Evol **8**:398-414.

21. **Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jagels K, Krogh A, McLean J, Moule S, Murphy L, Oliver K, Osborne J, Quail MA, Rajandream M-A, Rogers J, Rutter S, Seeger K, Skelton J, Squares S, Sulston JE, Taylor K, Whitehead S, Barrell BG.** 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature **393**:537-544.

22. **Fleischmann RD, Alland D, Eisen JA, Carpenter L, White O, Peterson J, DeBoy R, Dodson R, Gwinn M, Haft D, Hickey E, Kolonay JF, Nelson WC, Umayam LA, Ermolaeva M, Salzberg SL, Delcher A, Utterback T, Weidman J, Khouri H, Gill J, Mikula A, Bishai W, Jacobs WR, Venter JC, Fraser CM**. 2002. Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. J Bacteriol **184**:5479-5490.

23. **Zhang Y, Chen C, Liu J, Deng H, Pan A, Zhang L, Zhao X, Huang M, Lu B, Dong H, Du P, Chen W, Wan K.** 2011. Complete genome sequences of *Mycobacterium tuberculosis* strains CCDC5079 and CCDC5080, which belong to the Beijing family. J Bacteriol **193**:5591-5592.

24. **Bentley SD, Comas I, Bryant JM, Walker D, Smith NH, Harris SR, Thurston S, Gagneux S, Wood J, Antonio M, Quail MA, Gehre F, Adegbola RA, Parkhill J, de Jong BC.** 2012. The genome of *Mycobacterium africanum* West African 2 reveals a lineage-specific locus and genome erosion common to the *M. tuberculosis* complex. PLoS Negl Trop Dis **6**(2):e1552. doi:10.1371/journal.pntd.0001552.

25. **Garnier T, Eiglmeier K, Camus J-C, Medina N, Mansoor H, Pryor M, Duthoy S, Grondin S, Lacroix C, Monsempe C, Simon S, Harris B, Atkin R, Doggett J, Mayes R, Keating L, Wheeler PR, Parkhill J, Barrell BG, Cole ST, Gordon SV, Hewinson RG.** 2003. The complete genome sequence of *Mycobacterium bovis*. Proc Natl Acad Sci U S A **100**:7877-7882.

26. **Supply P, Marceau M, Mangenot S, Roche D, Rouanet C, Khanna V, Majlessi L, Criscuolo A, Tap J, Pawlik A, Fiette L, Orgeur M, Fabre M, Parmentier C, Frigui W, Simeone R, Boritsch EC, Debrie A-S, Willery E, Walker D, Quail MA, Ma L,**

Bouchier C, Salvignol G, Sayes F, Cascioferro A, Seemann T, Barbe V, Locht C, Gutierrez M-C, Leclerc C, Bentley S, Stinear TP, Brisse S, Medigue C, Parkhill J, Cruveiller S, Brosch R. 2013. Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of *Mycobacterium tuberculosis*. Nat Genet **45**:172-179.

27. **Wang J, McIntosh F, Radomski N, Dewar K, Simeone R, Enninga J, Brosch R, Rocha EP, Veyrier FJ, Behr MA.** 2015. Insights on the emergence of *Mycobacterium tuberculosis* from the analysis of *Mycobacterium kansasii*. Genome Biol Evol **7**:856-870.
28. **Richter M, Rosselló-Móra R.** 2009 Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A **106**(45):19126-19131.
29. **Delcher AL, Phillippy A, Carlton J, Salzberg SL**. 2002. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res **30**(11):2478-2483.
30. **Chan JZM, Halachev MR, Loman NJ, Constantinidou C, Pallen JM.** 2012. Defining bacterial species in the genomic era: insights from the genus *Acinetobacter*. BMC Microbiol **12**:302.
31. **Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM.** 2007. DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol **57**(1):81–91.

8.3 Additional unpublished analyses

8.3.1 Mixed species infection

One isolate from outside the outbreak village was classified as mixed infection based on BLASTn (151). In this isolate, both *M. tuberculosis* (4,204/10,004 reads) and *Mycobacterium avium* (5,551/10,004 reads) were detected. Therefore, to avoid bias in depth of coverage calculations, this isolate was excluded from the current analysis.

8.3.2 Use of an alternative SNP calling algorithm

To assess the impact of using a different SNP calling algorithm, I repeated SNP calling against the *M. tuberculosis* CDC1551 reference genome, using SAMtools' mpileup command and bcftools call. SNPs were filtered for quality using the same parameters as Unified Genotyper. Quality depth and the Mapping Quality RankSum were not calculated by SAMTools, and thus were not used for filtering. SNPs were filtered for strand bias using VCFtools –annotate- (147). SNPs identified by both callers are compared below.

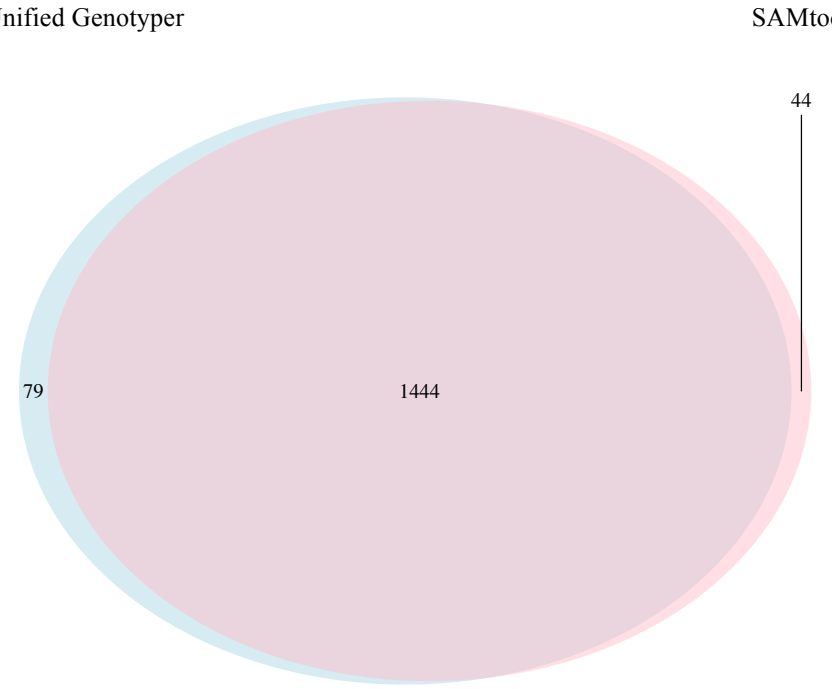


FIGURE 8-2 Venn Diagram of SNP loci
92.2% of SNP loci were identified by both callers. 5.0% were identified by Unified Genotyper alone, and 2.8% were identified by SAMtools alone.

To assess these minor differences would influence phylogenetic trees, I then produced the following maximum likelihood tree based on the SNPs identified by SAMtools, using the Tamura 3-parameter (160) model of nucleotide substitution:

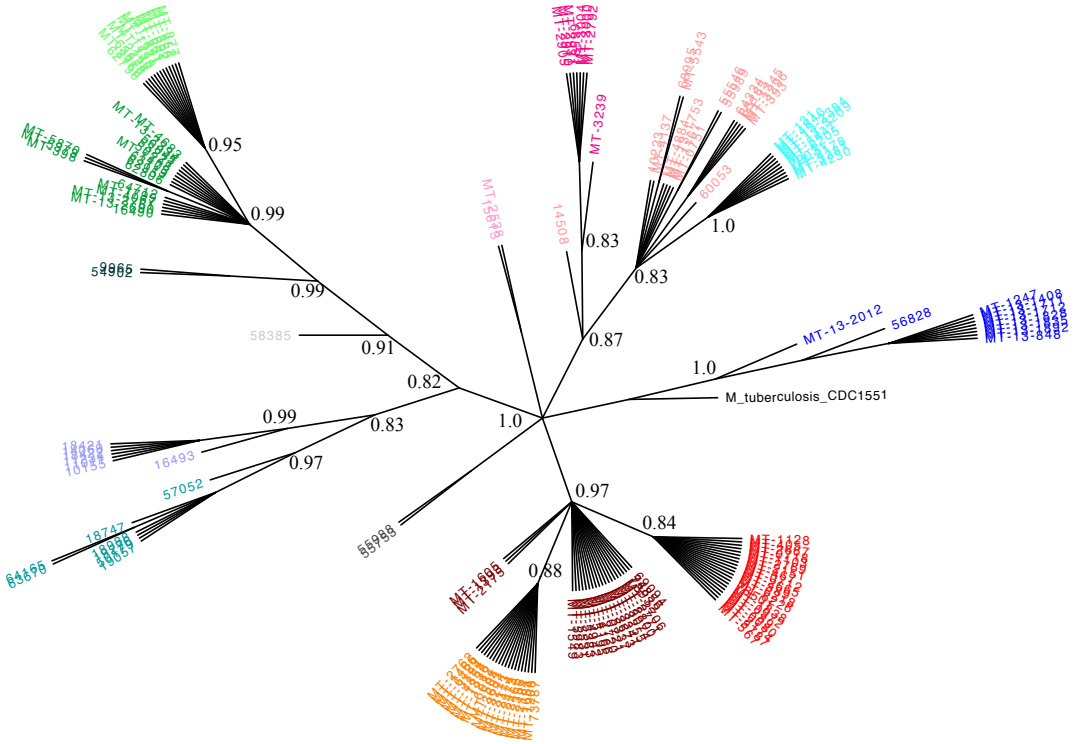


FIGURE 8-3 Maximum likelihood tree based on SNPs identified using SAMtools
SNPs were called using SAMtools mpileup and bcftools call, compared to the CDC155 reference genome. Isolates are coloured according to clusters identified with Unified Genotyper. Bootstrap proportions are indicated, starting proximally until the cluster-defining nodes. Branches are collapsed at 80% bootstrap threshold, and therefore branch length does not correspond directly to genetic distance.

All isolates are clustered in the same groups as they were previously, with SNPs called by Unified Genotyper, indicating that the phylogeny is robust to use of a different SNP calling algorithm.

CHAPTER 9. OBJECTIVE 5 – Manuscript V

Lee RS and Behr MA. The implications of whole genome sequencing in the control of tuberculosis. 2016. *Ther Adv Infect Dis*;3(2):47-62.

9.1 Preamble

During the course of this thesis work, WGS has gone from being applied to a single outbreak (3) to the gold standard for epidemiologic studies of tuberculosis, as well as many other infectious diseases (e.g., (145, 176-181)). Thus far, its use has been restricted predominantly to the research domain, though several regional public health departments and countries are currently implementing it as part their routine TB surveillance (e.g., British Columbia, Canada; the United Kingdom (UK), among others). As the use of WGS is increasing, many have proposed a new role for this tool in TB diagnostics, for both the detection of TB disease and prediction of drug resistance.

The following manuscript is an invited review that discusses the current status of WGS and its potential utility in the realm of diagnostic microbiology. While the focus of this replace is the application of WGS to clinical medicine, it is important to note that, should WGS replace conventional diagnostics, this has substantial implications for all epidemiologic studies of tuberculosis.

The reprint of this manuscript can be found in **Appendix 6**.

9.2 MANUSCRIPT V

The implications of whole genome sequencing in the control of tuberculosis

Robyn S. Lee [a,b,c,] Marcel A. Behr [b,c,1]

1. Corresponding author

Author affiliations:

a. McGill University, Department of Epidemiology, Biostatistics and Occupational Health

b. The Research Institute of the McGill University Health Centre

c. McGill International TB Centre

Corresponding author:

Dr. Marcel A. Behr

McGill International TB Centre

1001 boul Décarie, Block E, Mail Drop Point #EM33211,

Montréal, QC H4A 3J1 Canada

Phone: 514 934-1934, marcel.behr@mcgill.ca

Key words: *Mycobacterium tuberculosis*; whole genome sequencing; clinical microbiology; diagnostics; drug-resistance

Short title: Whole genome sequencing and tuberculosis

Abstract

The availability of whole-genome sequencing (WGS) as a tool for the diagnosis and clinical management of tuberculosis (TB) offers considerable promise in the fight against this stubborn epidemic. However, like other new technologies, the best application of WGS remains to be determined, for both conceptual and technical reasons. In this review, we consider the potential value of WGS in the clinical laboratory for the detection of *Mycobacterium tuberculosis* and the prediction of antibiotic resistance. We also discuss issues pertaining to data generation, interpretation and dissemination, given that WGS has to date been generally performed in research labs where results are not necessarily packaged in a clinician-friendly format. Although WGS is far more accessible now than it was in the past, the transition from a research tool to study TB into a clinical test to manage this disease may require further fine-tuning. Improvements will likely come through iterative efforts that involve both the laboratories ready to move TB into the genomic era and the front-line clinical/public health staff who will be interpreting the results to inform management decisions.

Introduction

Owing to advances in technology and reductions in cost, whole-genome sequencing (WGS) has been transformed from a centralized service used by a select few to interrogate single genomes into a relatively decentralized lab technique used by many to detect and track infectious pathogens [Long *et al*. 2014; Price *et al*. 2014; SenGupta *et al*. 2014; Snitkin *et al*. 2012; Quick *et al*. 2014, 2015]. This transformation has not spared the mycobacterial genus, with a number of papers presenting its application to the characterization of *Mycobacterium tuberculosis* cases and outbreaks [Walker *et al*. 2013; Bryant *et al*. 2013; Gardy *et al*. 2011; Lee *et al*. 2015; Casali *et al*. 2014; Jamieson *et al*. 2014b; Stucki *et al*. 2015; Roetzer *et al*. 2013; Guerra-Assuncao *et al.* 2015]. In this review, we will consider the opportunities presented by WGS for clinical management of tuberculosis (TB) across two conceptual spaces: diagnosis (*M. tuberculosis* detection) and treatment (prediction of antibiotic resistance). We recognize that the greatest utility for WGS will likely lie in countries with the highest TB burdens; however, as WGS requires substantial financial and technical infrastructure, we have situated this review in the setting of a high-resource country where this method may be more imminently implemented.

A brief description of WGS

WGS begins at the bench, with the extraction and purification of genomic DNA. In very brief detail, this DNA is typically fragmented into shorter pieces, which are then sequenced in 'reads' of 100-500 base pairs (bp) for bench-top sequencers. There are a number of different sequencing platforms available [Loman *et al*. 2012a; Kwong *et al*. 2015; Heather and Chan, 2015]. The choice of platform depends largely on the question, which in turn is dictated by clinical needs. If the aim is to identify unknown organisms or to characterize a novel bacterium, one might prefer a sequencer that generates longer reads (such as the PacBio RS by Pacific Biosciences, Menlo Park, CA, USA), as such reads enable more accurate de novo assembly [Loman *et al*. 2012a]. If the goal is to speciate the microorganism, determine drug resistance or resolve transmission networks, sequencers producing short reads can be used. Among the benchtop sequencers generating short read data, the most accurate platform currently available is the Illumina MiSeq (Illumina, San Diego, CA, USA) [Loman *et al*. 2012b] (though whether the difference in accuracy compared with another platform, the Ion Torrent PGM from ThermoFisher Scientific, Waltham, MA, USA, ultimately affects clinical inferences has been questioned [Harris *et al*.

2013]). In the analysis of such short read data, a reference-based approach is preferred [Loman *et al*. 2012a], wherein these reads are aligned ('mapped') to a reference genome. This is ideal for analysis of *M. tuberculosis*, given the absence of horizontal gene transfer in this species and the existence of complete, well-annotated reference genomes. Such a workflow for *M. tuberculosis* is illustrated in **Figure 9-1**.

With the Illumina MiSeq platform, short reads of up to 300 bps in length are produced. To identify the microorganism in question based on these reads, a variety of tools can be utilized. The Basic Local Alignment Search Tool (BLAST [Altschul *et al*. 1990]) compares reads with existing microbial DNA databases and uses an algorithm to identify the most likely microorganism. Other methods include classifying the microorganism based on how well reads align to conserved coding sequences within phyla or species ('clade-specific marker sequences' [Segata *et al*. 2012]) or k-mer-based approaches [Wood and Salzberg, 2014]. In the latter, reads are divided into segments of k bases in length (called 'k-mers') that are compared with a database of known k-mer sequences from selected microorganisms. The best identification is determined as the microorganism with the highest proportion of matching k-mers.

Once reads have been assigned the identity '*M. tuberculosis*', they are subsequently mapped to the corresponding sequence on the reference genome to identify differences (i.e. variants) in the sample compared with this reference. There are several key considerations when performing such reference-based analyses. First, the choice of an appropriate reference genome is crucial; if the reference is too dissimilar from the isolate in question, large numbers of reads will not be mapped and these data (and all variation therein) will be ignored. Second, alignment to GC-rich repetitive regions can be difficult, as reads may map to more than one location, thereby producing inconclusive matches. Such regions include the PE-PPE family proteins, which comprise ~10% of the coding sequence of *M. tuberculosis* [Cole *et al*. 1998]. To reduce the risk of false- positive results, the PE-PPE regions and mobile elements are typically excluded from analyses [Comas *et al*. 2010; Roetzer *et al*. 2013]. Alternatively, one could perform targeted sequencing using a platform capable of generating longer reads that span repetitive regions. However, this would incur additional expense, as well as technical/bioinformatics requirements, and may not provide additional information of use for clinical applications.

Using a reference-based approach, single nucleotide polymorphisms (SNPs; i.e. a difference in a single base in the genome compared to the reference) and insertions/deletions (indels) present in the test isolate can be identified ('called') compared with the referent. This process, the quality control steps therein and the different tools used for identifying SNPs are reviewed in detail elsewhere in [Pabinger *et al*. 2014; Olson *et al*. 2015]. For the purposes of this work, we have focused on the utility of WGS for the clinician and, in particular, the use of these SNPs to predict drug resistance. In *M. tuberculosis* research, SNPs have also been used to extensively to delineate transmission networks, however, an in-depth discussion of this utility is beyond the scope of this review. The interested reader is directed to the several examples in the literature of its use in TB outbreak investigations [Gardy *et al*. 2011; Stucki *et al*. 2015; Lee *et al*. 2015; Torok *et al*. 2013; Kato-Maeda *et al*. 2013; Schurch *et al*. 2010; Ocheretina *et al*. 2015; Walker *et al*. 2013; Roetzer *et al*. 2013]. It is worth noting at this point that genotyping is occasionally required for clinical care, for instance, to rule out laboratory cross-contamination as a false-positive cause of a positive culture, or when trying to determine when a TB recurrence is due to relapse of the original infection versus exogenous reinfection. For both of these applications, the lessons of outbreak investigation indicate that WGS has higher resolution than traditional typing methods, such as spoligotyping, mycobacterial interspersed repetitive units (MIRUs), or restriction fragment length polymorphism (RFLP) [Gardy *et al*. 2011; Lee *et al*. 2015; Walker *et al*. 2013; Roetzer *et al*. 2013]. Therefore, it can be inferred that, for both situations, if the traditional method returns a result of 'different strain', WGS is likely not necessary to answer the clinical question. If, however, the traditional typing method returns a matched pattern, WGS may be required to confidently distinguish a related strain due to ancestry from a true match, with the latter being observed during laboratory cross-contamination or relapse.

Regardless of the application, the quality of WGS data depends on a number of factors, including the desired length of the sequencing reads and the cycle time [Quick *et al*. 2015]. These parameters in turn affect the turnaround time for results. Considering the most frequently used benchtop sequencers, raw sequencing results can be available in a clinically attractive span of just a few hours (for the Ion Torrent PGM) to as much as 39 hours with MiSeq for paired end 250 bp reads. By adjusting the sequencing protocol for MiSeq, it may be feasible to reduce this

time frame without affecting key inferences, such as species and strain assignments [Quick *et al.* 2015]. An important consideration when making such adjustments is the 'depth of coverage'; the more reads that span a position in the reference genome, the more support there is for the base identified. The optimal depth of coverage to detect clinically relevant variants needs to be determined.

Another factor influencing the time to obtain these data is whether samples are batched or run independently. According to Quick and colleagues [Quick *et al*. 2015], the MiSeq can sequence up to ~100 isolates simultaneously. In our experience, the MiSeq 250 bp paired-end sequencing can generate a minimum of 10 million reads; if 20x coverage is desired, only ~57 isolates of *M. tuberculosis* can be run simultaneously [Lander and Waterman, 1988]. A batched approach such as this is typical in research labs and is clearly less expensive on a per-unit basis, as running a single isolate would cost the same as the whole collection of samples. Unfortunately, waiting until a queue of specimens has accumulated is not ideal for clinical labs, which need to process samples immediately on arrival and send reports 24 hours a day. A newer method, the Nanopore MinIon (Oxford Nanopore Technologies, Oxford, UK), offers much promise in addressing this problem. The MinIon runs a single sample at a time and was able to correctly speciate two *Salmonella enterica* isolates as well as place them in epidemiologic context within 2h [Quick *et al*. 2015]. Earlier diagnosis and detection of SNPs connoting drug resistance could allow for more rapid initiation of treatment, compared with waiting for results from a batched analysis. However, the advantage of rapid results offered by the MinIon is currently offset by high error rates as reported by [Laver *et al*. 2015; Mikheyev and Tin, 2014; Quick *et al*. 2015]. While sequencing chemistry is improving and bioinformatics approaches are being developed to increase accuracy [Jain *et al.* 2015], further studies are needed to evaluate this method. As of yet, the MinIon has not been utilized for *M. tuberculosis*. It might be that these different platforms offer complementary opportunities for the clinical lab, for instance by using the Nanopore technology to rapidly speciate pathogenic organisms and the MiSeq for ongoing epidemiologic surveillance.

WGS for detection of *M. tuberculosis*, including the prediction of drug resistance

In the clinical mycobacteriology lab, the goal is to secure a diagnosis of active TB and to

provide clinicians with guidance on which antibiotics they should or should not prescribe for their patients. These two goals have classically been achieved with phenotypic tests, some dating to the 19th century. This begs the obvious question of whether WGS can help modernize the TB lab, with the goal of offering faster and more accurate results.

The current clinical workflow for detection of *M. tuberculosis* in Canada is illustrated in **Figure 9-2**. Variations of this pathway may be seen in comparable high-resource countries. For more detailed reviews of *M. tuberculosis* laboratory diagnosis, the reader is referred to the literature [Parrish and Carroll, 2008, 2011; Drobniewski *et al*. 2013; Noor *et al*. 2015]. In brief, specimens from TB suspects are sent for smear microscopy to ascertain the presence of acid-fast bacilli. This test identifies the most infectious patients (i.e. with 'smear-positive' disease) [Behr *et al.* 1999]. Results of smear microscopy should be available within 24h of receipt [Parrish and Carroll, 2011], however this method has low sensitivity [Steingart *et al*. 2006a, 2006b] and cannot distinguish *M. tuberculosis* from non-tuberculous mycobacterium. Regardless of the results of microscopic examination, the same specimens are processed for culture, as detailed by Parrish and Carroll [Parrish and Carroll, 2011]. The culture is usually done using both solid and liquid media (typically mycobacterial growth indicator tubes [MGITs]), with growth usually observed in 1-3 weeks, depending on the mycobacterial inoculum in the sample [Chihota *et al*. 2010; Fadzilah *et al*. 2009]. Once growth is observed (on solid media) or flagged by the machine (in the case of MGITs), a positive culture can be assigned a presumptive identification as *M. tuberculosis* complex using a DNA probe, usually within 24h [Ichiyama *et al*. 1997]. Cultures are then sent to a reference laboratory for formal species confirmation and for drug susceptibility testing (DST) by phenotypic (i.e. growth-based) assays.

Superimposed on this classic workflow (smear microscopy, culture, then DST), laboratories have overlaid molecular testing over the past two decades, using a variety of different platforms and clinical strategies. The first molecular tests approved were only licensed for the speciation of smear microscopy-positive samples [Parrish and Carroll, 2011], so their key role was in assigning a microbial name to such a sputum sample [Vuorinen *et al*. 1995; Carpentier *et al*. 1995]. Then, with time and experience, it became recognized that nucleic acid amplification testing could be offered on smear-negative samples where there was a high clinical suspicion of

TB [Centers for Disease Control and Prevention (CDC) 2009]. To reduce costs of controls, these 'rapid' first generation tests were generally batched and as a result, might only have been done twice or three times per week, depending on laboratory volume. More recently, the GeneXpert (Cepheid Inc., Sunnydale, CA, USA) has offered a random-access real-time nucleic acid amplification test, which can be done on a single sample, without having to wait for samples from other patients. GeneXpert is conducted directly on the clinical specimen to detect both the presence of *M. tuberculosis* DNA and mutations in the *rpoB* gene that predict resistance to the first-line drug, rifampin. In principle, results can be available in under 2h [Boehme *et al*. 2010]. In practice, turnaround time depends on logistics; most testing is done in laboratories rather than clinics, necessitating delays due to shipping and handling [Alvarez *et al*. 2015]. The specificity of GeneXpert for *M. tuberculosis* detection is high, reported at >98%, but the sensitivity varies by smear status [Boehme *et al*. 2010; Steingart *et al*. 2014; Sohn *et al*. 2014], site (e.g. respiratory versus extrapulmonary) and type of sample (e.g. lymph node versus pleural [Maynard-Smith *et al*. 2014; Denkinger *et al.* 2014]). While GeneXpert is currently the fastest and arguably most useful diagnostic test in many parts of the world, it may be that its enduring legacy is catalyzing a paradigm shift away from phenotypic testing, towards genetic detection of *M. tuberculosis* as the primary goal of the TB lab. If true, then the same pre-analytic principles (collecting sputum, delivering to lab, rendering the sample safe, extracting DNA) can serve as the basis for a more comprehensive interrogation of the mycobacterial genome, going beyond the *rpoB* gene to characterize the complete genome of the causative organism.

WGS for diagnosis

Until recently, the utility of WGS for de novo diagnosis of *M. tuberculosis* was unclear. WGS had relied exclusively on enriched DNA obtained from a pure bacterial culture, at which point the patient would have already been diagnosed. More recently, studies have examined the feasibility of sequencing *M. tuberculosis* directly from the clinical specimen [Doughty *et al*. 2014; Brown *et al*. 2015]. Sequencing eight smear positive samples, Doughty and colleagues obtained only 0.002x to 0.7x depth of coverage, with 20-99% of reads sequenced mapping to the human genome rather than *M. tuberculosis* [Doughty *et al*. 2014]. Brown and colleagues obtained similar results when sequencing directly from clinical samples, but when an oligonucleotide enrichment protocol was applied, they were able to obtain at least 20x depth of

coverage on 20/24 smear positive, culture positive isolates, providing sufficient sequence depth to confidently speciate the organism present [Brown *et al*. 2015].

If WGS is to be applied on the patient sample, the conceptual advantage is a more rapid result. However, the vast majority of samples are negative for *M. tuberculosis*, even in a high-incidence setting [Demers *et al*. 2012], so some form of triage is needed to select the samples most likely to benefit from direct WGS. Furthermore, sputum is contaminated with host and other bacterial DNA, complicating bioinformatics analyses and reducing the overall depth of coverage obtained for the *M. tuberculosis* genome [Doughty *et al*. 2014]. While low coverage may not preclude the ability to confidently detect *M. tuberculosis*, it could seriously undermine the capacity to detect mutations associated with drug resistance (as shown by Doughty and colleagues [Doughty *et al*. 2014]), where the greatest clinical value of WGS may lie. In sum, these studies provide proof-of-principle that WGS of *M. tuberculosis* directly from clinical specimens is feasible, but the cost of the enrichment protocol (USD$350 per sample), the requirement for technical expertise and equipment, and the need for real-time bioinformatics to convert sequence files into clinically meaningful lab reports all present challenges to WGS supplanting smear microscopy and nucleic acid amplification as the primary test performed on clinical specimens.

If instead WGS is applied on the positive culture, then the benefit of rapidity has been lost, as the patient should already be isolated and started on treatment, based on either smear microscopy, a nucleic acid amplification test or the Accuprobe result on the culture. In this case, WGS may offer a different opportunity, which is a more rapid identification of antibiotic resistance.

WGS for resistance

In 2013, 3.5% of incident TB cases worldwide (95% confidence interval [CI] 2.2-4.7%) were estimated to have multidrug-resistant (MDR) TB, with an enrichment to 20.5% in cases with previous treatment (95% CI 13.6-27.5%) [World Health Organization, 2015]. As there is no evidence for ongoing acquisition of foreign DNA by *M. tuberculosis*, resistance occurs due to mutations in the chromosomal DNA, some of which have been mapped and mechanistically linked to the resistance phenotype [Nebenzahl-Guimaraes *et al*. 2014]. Phenotypic testing of a positive culture (called indirect DST) is the current gold standard for *M. tuberculosis*. The need

for level 3 containment facilities and the requirement to perform an appropriate number of tests to maintain competence, however, have conspired to direct this most clinically meaningful assay to reference labs, entailing delays due to transport and handling. Therefore, while it is stated that first-line susceptibility results can be obtained in 2-4 weeks [Perkins and Cunningham, 2007; Migliori *et al*. 2008], such estimates reflect the time for work to be performed in the reference lab. When considering the time from sample acquisition to a final report, others provide longer timeframes, up to 2 months [Parrish and Carroll, 2008]. Until this information is available, the clinician faces an immediate dilemma, which is: 'What do I prescribe now?'. Inappropriate treatment risks generating further drug resistance, but delaying treatment until a final report is provided risks deleterious treatment outcomes [Park *et al*. 1996]. While one option is to attempt phenotypic testing directly on the patient sample (called 'direct DST'), there are still delays with the time to obtaining cultures, and susceptibility testing on the sputum sample brings its own challenges, since it is difficult to standardize the inoculum for such assays. It is at this moment of indecision that a molecular test could provide the most immediate clinical guidance, as exemplified by the GeneXpert test. For examples of molecular tests, along with sensitivity and specificity for respective drugs, see **Table 9-1**.

As most rifampin-resistant isolates are also isoniazid-resistant, the GeneXpert uses *rpoB* mutations associated with rifampin-resistance as a proxy for multi-drug resistance. However, not all rifampin-resistant organisms are isoniazid- resistant (i.e. there can be rifampin mono-resistance) and indeed, not all isolates predicted to be rifampin-resistant are confirmed on phenotype-based testing [Steingart *et al*. 2014]. In addition, not all rifampin-resistant isolates are detected based on the currently assessed mutations [Sanchez-Padilla *et al.* 2015; Jamieson *et al*. 2014a]. Finally, GeneXpert may fail to detect hetero-resistance, i.e. resistance-connoting mutations present in subpopulations within the patient [Zetola *et al*. 2014]. For all of these reasons, a broader-based assay, such as WGS, could offer the greatest clinical utility at this point in the diagnostic process, by looking beyond the targets of the current molecular assays.

By sequencing the whole genome, in theory all resistance-connoting mutations that can guide clinical treatment can be identified by comparing the genome of the patient isolate with detailed databases of known resistance markers [Sandgren *et al*. 2009; Flandrois *et al*. 2014]. In practice,

this will work, if (a) these markers accurately predict in vitro phenotypic resistance, and (b) these markers predict clinical outcome. For the latter, we are unaware of studies that have directly assessed the utility of WGS data for predicting patient response to treatment. For the proximal goal of linking WGS to phenotypic resistance, there are emerging data which present a mixed message. Using online databases, supplemented with an updated search of the literature, Coll and colleagues [Coll *et al*. 2015] developed a mutation library and examined the concordance between genotypic predictions and phenotypic data for 788 isolates from diverse geographic settings. Among the drugs with sufficient phenotypic data (rifampin (RIF), isoniazid (INH), ethambutol (EMB), pyrazinamide (PZA) and streptomycin (STR)) as well as second-line drugs (amikacin (AMK), capreomycin (CAP), ethionamide (ETH), kanamycin (KAN), moxifloxicin (MOX), ofloxacin (OFX)), the sensitivity of WGS for predicting resistance was highest for INH and RIF at 92.8% (95% CI 89.9-95.7) and 96.2 (95% CI 93.9-98.5). At the other end of the spectrum, the sensitivity of WGS for PZA resistance was only 70.9% (95% CI 62.4-79.4). Thus, if WGS replaced phenotypic testing, one-twelfth of INH-resistant and one-third of PZA-resistant cases would receive these potentially hepatotoxic drugs, with little or no benefit. Specificity of WGS was highest for INH and RIF at 100% (95% CI 100-100%) and 98.1% (95% CI 96.8-99.4%), respectively, but for other drugs, specificity was as low as 81.7% (EMB).

In the same manuscript [Coll *et al*. 2015], Coll and colleagues also compared the performance of their database with KvarQ, a software that uses pre-specified 'testsuites' of known resistance-connoting mutations and other regions of interest to predict resistance [Steiner *et al*. 2014]. Using phenotypic data as the gold standard, sensitivity was substantially lower for nearly all drugs using the KvarQ method (though 95% CIs overlapped for all except EMB and KAN). Among first-line drugs, only RIF yielded similar point estimates to those obtained with Coll and colleagues' mutation library, with sensitivity of 95.8% (95% CI 93.4-98.2%), while sensitivity for INH was only 86.9% (95% CI 83.1-90.7%). No results were available for ETH and CAP using the KvarQ software. Specificity was generally higher using KvarQ, though this difference was only significant for EMB and STR. Specificity for RIF was similar to that obtained with the mutation database, at 97.9% (95% CI 96.5-99.3%).

In a similar study [Walker *et al.* 2015], Walker and colleagues selected 23 candidate resistance-

associated genes from the literature [Sandgren *et al*. 2009] and then used an algorithm to characterize mutations (SNPs and indels) within these genes and their promoter regions as resistance-connoting or benign. In a training dataset of 2099 isolates, 120 resistance-connoting mutations were identified, 772 were classified as benign and 101 could not be classified as either (called 'uncharacterized'). The resistance-connoting and benign mutations identified in this training dataset were then used in a validation study on an additional 1552 genomes, 29% of which were resistant to at least one drug on drug susceptibility testing (DST). Using these mutations, authors were able to predict 89.2% of phenotypes as resistant or susceptible. 10.8% of phenotypes could not be predicted, as these contained mutations that had not been characterized. Among those where phenotype could be predicted and considering predictions for each drug independently, 112 of 6892 with drug-sensitive DST were predicted to be resistant based on WGS (1.6%), while 94 of 1221 with drug-resistant DST were erroneously predicted to be drug-sensitive (7.7%). The latter may be due to mutations with unknown function outside the 23 candidate genes interrogated. This is similar to Farhat and colleagues [Farhat *et al*. 2013]; in this study, authors performed targeted deep sequencing of known resistance genes to verify that resistance mutations were absent in subpopulations within isolates. They found that 13/47 isolates with phenotypic resistance had no previously known mutations. Unexplained resistance, wherein phenotypic resistance is present but known resistance-connoting mutations are absent has been most pronounced for PZA [Hewlett *et al*. 1995] and second-line drugs. For example, Farhat and colleagues [Farhat *et al*. 2013] found that, among isolates resistant to ciprofloxacin, KAN and CAP, 2/3, 6/18 and 1/ 6 isolates, respectively, had unexplained resistance. As the reliability of phenotypic testing is least well established for these drugs [Horne *et al*. 2013], this is where there is the greatest need for WGS, but presently also the greatest knowledge gap.

In clinical medicine, the physician wants to know whether the isolate has a resistance-connoting mutation or not, so that treatment can be tailored accordingly. Indeterminate test results offer little clinical guidance, and often steer clinicians to other antibiotics, where feasible. While it is logical to exclude isolates with uncharacterized mutations from a scientific paper that aims to understand resistance, in a clinical laboratory, these have to be reported one way or the other. Analyses that classified such uncharacterized mutations as predictive of phenotypic susceptibility greatly affected test parameters; the sensitivity of WGS for INH and RIF resistance dropped

from 94.2% (95% CI 91.1-96.5%) and 96.8% (95% CI 94.1-98.5%) with uncharacterized mutations excluded to 85.2% (95% CI 81.1-88.7%) and 91.7% (95% CI 87.9-94.5%) with uncharacterized mutations included, respectively. Sensitivity for PZA resistance in the latter analysis was the lowest overall, at only 24% (95% CI 17.9-30.9%). Until such mutations can be confidently assigned to the appropriate phenotype, it would seem that parallel, or at the least, sequential phenotypic testing should remain part of the diagnostic pathway.

Furthermore, these publications generally included biased samples, with relatively high proportions of drug-resistant isolates. As many clinical labs identify primarily drug-sensitive isolates, the operating parameters of WGS for this purpose may change when evaluated against more representative samples. While authors had generally high specificity for most drugs, the predictive value depends on the underlying prevalence of drug resistance. In a country such as Canada, which detected RIF resistance among only 17 of 1380 *M. tuberculosis* complex isolates analyzed in 2013 [Public Health Agency of Canada, 2015], a specificity of 98.1-99.2% and sensitivity of 91.7-96.2% based on the results of Coll and colleagues [Coll *et al*. 2015] and Walker and coworkers [Walker *et al.* 2015] would equate to ~18 false positives per year, with a positive predictive value of only ~46%. Without subsequent phenotypic testing, these cases would be subject to second-line treatment, with prolonged, unnecessary hospitalization. Thus, WGS may be best reserved only for individuals in which there was a higher pretest probability of resistance (based on some a priori criteria for the use of WGS, e.g. previous treatment).

Despite these limitations, it is clear that WGS offers magnitudes more information than the molecular methods listed in **Table 9-1**, with the potential of greatly advancing clinical diagnostics for *M. tuberculosis*. While the WGS database of Coll and colleagues [Coll *et al*. 2015] performed similarly to GeneXpert for RIF resistance, it also allowed for determination of INH mutations, and had an overall accuracy of 95.8%, as compared to 93.1% for MTBDRplus (Hain Lifescience, Nehren, DE) ($p<0.0004$). Accuracy was also higher for second-line drugs compared with MTBDRsl (Hain Lifescience, Nehren, DE) (96.3% versus 93.7%, $p<0.0047$). Walker and colleagues [Walker *et al*. 2015] showed similar sensitivity and specificity of their algorithm for determining the correct phenotype using WGS as the collective results of MTBDRplus, MTBDRsl and AID (AID Diagnostika, Strassberg, DE) line probe assays (LPAs).

In addition, while synonymous SNPs can present as false positives on both LPA or GeneXpert, Walker and colleagues were able to classify these as benign.

Overall, these data support the great potential of WGS as a tool to predict resistance. However, databases of *M. tuberculosis* genomes, along with associated phenotypic data, are essential to identify unrecognized and emerging mutations. In addition, our ability to accurately predict phenotypic resistance is limited by our understanding of epistasis (the interaction between mutations, which can influence phenotype [Trauner *et al*. 2014]); mutations associated with resistance have been found in phenotypically sensitive bacteria [Walker *et al*. 2015], in some cases potentially due to interaction with other mutations in the genome. Until additional data are gathered, it can be foreseen that WGS may serve as an added, rather than a replacement test, on the diagnostic pipeline (**Figure 9-2**). This would incur added costs to the lab, something that is clearly less attractive than WGS simply replacing drug susceptibility testing (DST), with all its labor and reagent costs. One need look no further than the example of HIV treatment to imagine a world where genotype-based data are used to predict drug resistance, and hence treatment decisions. However, for all of the aforementioned reasons, we submit that reference labs need to maintain competence in phenotypic DST for the foreseeable future.

Another issue for clinical application of WGS is timeliness of reporting. As of yet, two papers reported on the application of WGS in 'real-time' to clinical cases: a case report of a patient [Koser *et al*. 2012] with extremely drug-resistant (XDR) TB (defined as MDR TB plus resistance to an injectable second-line drug and a fluoroquinolone) and a prospective cohort of patients in the United Kingdom suspected of having XDR TB [Witney *et al*. 2015]. Koser and colleagues successfully obtained sequence data from a 3-day-old MGIT culture, identifying two concurrent but distinct strains of *M. tuberculosis* [Koser *et al*. 2013]. Predicted resistance and sensitivity concurred with phenotypic results for all drugs tested, while WGS predicted resistance to an additional five drugs. While WGS results had no impact on treatment, WGS did identify a mutation in the gene activating para-aminosalicylic acid (PAS) in the minority strain, despite a phenotypic determination of PAS-sensitive. Unfortunately, the functional impact of this was unknown. Witney and colleagues [Witney *et al*. 2015] selectively applied WGS to six cases with potential XDR TB, identified over 6 years in London, with multiple isolates sequenced per

patient. Results for five out of six cases were available in a clinically actionable time frame. Genotypic and phenotypic resistance were 100% concordant for INH and RIH, while discrepancies were reported in PZA, EMB, fluoroquinolones (OFX and MOX), AMK, KAN, CAP, PRO and PAS. In terms of clinical utility, WGS data helped guide treatment decisions by confirming PZA resistance in one case, and refuting an XDR diagnosis in favor of MDR in another. For another case, clinicians decided to continue with treatment with EMB, despite development of phenotypic resistance, as WGS failed to identify mutations in *embA* or *embB* that could explain the change in DST.

The Witney and colleagues study also illustrated that for WGS data to be used clinically, the results need to be analyzed rapidly and presented in a clear, easily interpretable manner. Several groups have produced online tools (e.g. 'PhyResSE' [Feuerriegel *et al.* 2015] and 'TB Profiler' [Coll *et al.* 2015]) wherein raw sequencing data for an isolate can be uploaded and analyzed for resistance-connoting mutations. As mentioned previously, the KvarQ software can also predict resistance from raw sequencing data; in contrast to PhyResSE and TB Profiler, this can be done on a local server [Steiner *et al*. 2014]. Yet, despite efforts to make these reports accessible to the wider scientific community, knowledge of genomics and/or bioinformatics is still required to interpret results. As an example, the quality of SNPs is provided with details such as depth of coverage, a parameter that most clinicians would be uncomfortable judging. Presently, PhyResSE and TB profiler are explicitly for research purposes only, which poses regulatory hurdles to the delivery of results destined for the clinical chart. Witney and colleagues [Witney *et al.* 2015] piloted a WGS report during the course of their study, but, unfortunately, clinician perception of this report and its interpretability was not assessed. Furthermore, though 'best practices' have been proposed for identifying SNPs [Olson *et al*. 2015], the current bioinformatics workflows used to analyze WGS data remain largely unstandardized. For implementation in the clinical lab, appropriate quality control measures [Clinical and Laboratory Standards Institute, 2014] and a standardized workflow need to be established. The lessons of the past five decades of emerging antibiotic resistance have demonstrated that even a simple dichotomous test result, i.e. resistant or susceptible, does not always predict appropriate care. Therefore, the application of WGS-based results to clinical care may benefit from evaluations done by experts in implementation science, rather than genomics or microbiology.

Conclusion

Offering increased resolution and substantially more data compared with conventional methods, WGS has revolutionized the arena of molecular epidemiology. Now, it seems poised to do the same for the clinical microbiology laboratory. The appeal of WGS for *M. tuberculosis* (and other pathogens) lies in the quantity of data provided; with one test, an organism can be speciated, resistance mutations can be detected and the strain can be placed in the context of the local epidemiology. The challenge of WGS also lies in the quantity of data provided; the same test can occupy a team of bioinformaticians, yet generate results that few clinicians can currently interpret. Furthermore, for WGS data to be clinically useful, results must be available in sufficient time to guide patient care. Recent advances such as sequencing directly from clinical samples and the rapid workflow of the Nanopore MinIon may facilitate this. The decision to whom this 'test' will be applied is also critical. Though no studies to date have examined cost-effectiveness of implementing WGS, it can be predicted that application of this test to all, unselected samples without removing other steps in laboratory workflow could be prohibitively expensive. Therefore, it can be foreseen that WGS will be applied selectively, for instance, on patients with Rifampin resistance mutations detected by the GeneXpert assay.

The issues raised above are only further amplified when contemplating the countries of the world that suffer the greatest burden of TB and have the highest prevalence of drug-resistant strains. While it is clearly feasible to ship sequencing machines around the world, as has already been done with the GeneXpert platform, it is not as simple to distribute the technical and bioinformatic expertise required for next-generation sequencing where it is needed. A potential solution to the latter is open-source coding and online data treatment, but this is currently lacking for clinical use, even in settings with expertise in these methods. Ultimately, what is needed is an easy-to-use software complete with a graphical user interface that is capable of converting data-intense sequence files into a simple, concise clinical message. As done with GeneXpert [Theron *et al*. 2014b], these outputs then need to be field-tested in settings with a sufficient burden of drug-resistant TB to enable evaluation of whether test results altered treatment decisions and clinical outcomes. The relatively small number of MDR TB patients in countries such as Canada may preclude a formal evaluation of patient outcomes, simply due to sample size considerations.

In order to assess its clinical utility for resource-rich countries where its use has been pioneered, we may need to first embed WGS in treatment studies conducted in the developing world, where the challenge posed by TB and drug resistance remains the greatest.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: *M. tuberculosis* genomic epidemiology work in the laboratory of MAB is supported by the Canadian Institutes of Health Research (MOP# 125858).

Conflict of interest statement

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

FIGURES

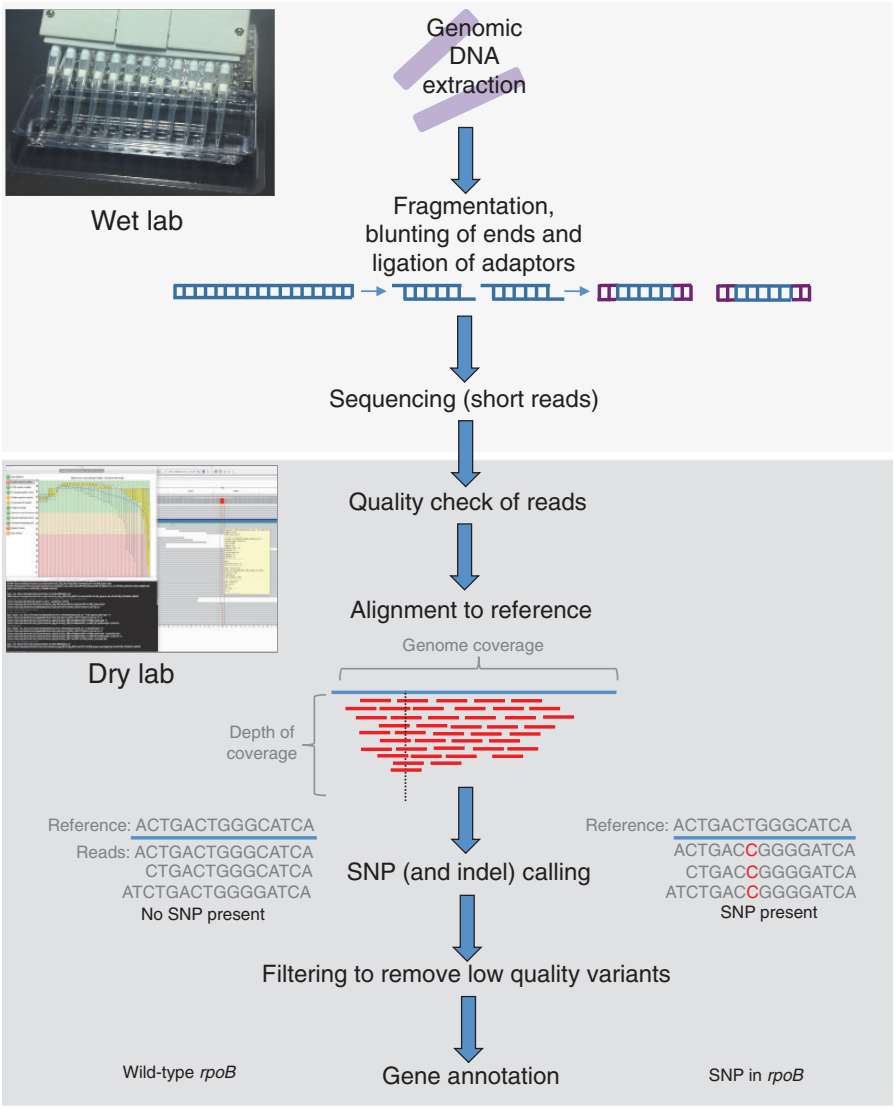


FIGURE 9-1. WGS workflow for *Mycobacterium* tuberculosis. In brief, whole-genome sequencing (WGS) begins in the wet lab (top panel), wherein genomic DNA (gDNA) is extracted. For a *M. tuberculosis* culture, this is done in a biosafety level 3 laboratory. After DNA extraction, library preparation is conducted, wherein genomic DNA is fragmented into pieces. Uneven ends of gDNA are blunted and adaptor sequences are added. After passing quality control, libraries are advanced to sequencing. Further analysis occurs in the dry lab (bottom panel). Potential contamination is assessed and the quality of sequencing is evaluated on a per isolate basis, including the examination of Phred quality scores of the sequenced bases (where $Phred=-10*\log P_{error}$). FastQC, for example, is a software that can be used for such quality control, and is applied directly on raw sequence data (available from http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/, shown in the screenshot). Adaptors (and potentially low-quality base pairs) are trimmed and reads of length under a prespecified limit (e.g. 70 base pairs used by the 1000 Genomes Project) may be excluded (not shown). High-quality reads are aligned to a reference genome (this can be visualized in Integrative Genomics

Viewer, also shown in screenshot [Thorvaldsdottir *et al*. 2013]), and metrics such as genome coverage (the percentage of the reference genome that has at least one read mapped to it) and depth of coverage (the average number of reads mapped to each locus) are evaluated. Isolates are retained if *a priori* quality measures are met. Reads are excluded if they map to more than one locus in the genome, and additional quality measures may be applied such as removing polymerase chain reaction duplicates and local realignment around indels. Once quality control steps are conducted, single-nucleotide polymorphisms and indels can then be 'called' compared with the reference genome. Low-quality variants are then removed using various filtering parameters to reduce the number of false positives. Genes are then annotated and repetitive regions and mobile elements may be filtered out of further analyses.

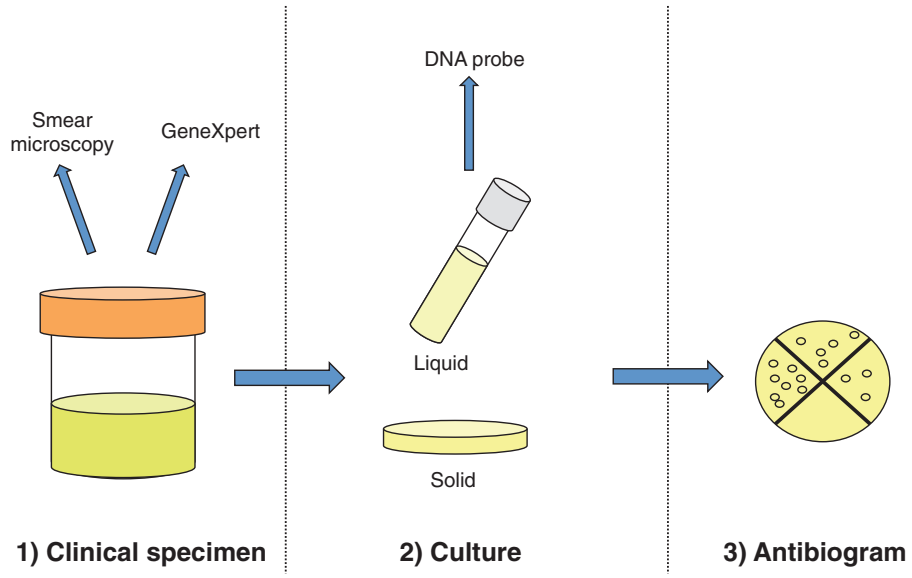


FIGURE 9-2. Clinical diagnostic workflow for *Mycobacterium* tuberculosis. The three main steps in the current diagnostic workflow for *M. tuberculosis* are shown. As described in the text, whole-genome sequencing may have a potential role at each of these steps: (1) by being applied directly to the unprocessed clinical specimen or (2) by being conducted on the positive culture to predict drug resistance.

TABLES

TABLE 9-1. Examples of molecular diagnostics for drug resistance in *M. tuberculosis*

Molecular test	Drug	Gene(s) targeted	Test performed on?	Sensitivity	Specificity	Turnaround time	Publication Type	Reference
GeneXpert (Cepheid Inc.)	Rifampin	*rpoB*	Raw clinical specimen (sputum + other respiratory specimens)	95 (95% CrI 90-97); range 33-100	98 (95% CrI 97-99); range 83-100	1h-4 d (run time <2h)	Meta-analysis	[Steingart KR *et al.* 2014]
MTBDR*plus* (Hain Lifesciences)	Rifampin	*rpoB*	Combined data for DNA from clinical specimen (respiratory+ non-respiratory samples) + purified DNA from culture	98.4 (95% CI 95.1-99.5); range 94-100*	98.9 (95% CI 96.8-99.7); range 95-100*	6h-2d*	Meta-analysis	[Ling *et al.* 2008]
	Isoniazid	*katG, inhA*	Combined data for DNA from clinical specimen (respiratory+ non-respiratory samples) + purified DNA from culture	88.7 (95% CI 82.4-92.8); range 57-100*	99.2 (95% CI 95.4-99.8); range 92-100*	6h-2d*	Meta-analysis	[Ling *et al.* 2008]
INNO-LiPA RifTB (Innogenetics)	Rifampin	*rpoB*	DNA from clinical specimen (including non-respiratory)	Range 80-100%	All 100%	Not reported	Meta-analysis	[Morgan *et al.* 2005]
			Purified DNA from culture	Range 82-100%	Range 92-100%	Not reported	Meta-analysis	[Morgan *et al.* 2005]

MTBDR*sl* (Hain Lifesciences)	Fluoroquinolones (including oxyfloxacin and levofloxacin)	*gyrA*	DNA from clinical specimen (smear positive sputum)	85.1 (95% CI 71.9-92.7); range 50-100	98.2 (96.8-99.0); range 91-100	8h-2d, 2 studies	Meta-analysis	[Theron *et al.* 2014a]
			Purified DNA from culture	83.1 (95% CI 78.7-86.7); range 57-100	97.7 (95% CI 94.3-99.1); range 77-100	1d (after 1st line)-10d, 2 studies	Meta-analysis	[Theron *et al.* 2014a]
	Aminoglycosides (including kanamycin, amikacin, capreomycin)	*rrs*	DNA from clinical specimen (smear positive sputum)	94.4 (95% CI 25.2-99.9); range 9-100	98.2 (95% CI 88.9-99.7); range 67-100	8h-2d, 2 studies	Meta-analysis	[Theron, *et al.* 2014a]
			Purified DNA from culture	76.9 (95% CI 61.1-87.6); range 25-100	99.5 (95% CI 97.1-99.9); range 86-100	1d (after 1st line)-10d, 2 studies	Meta-analysis	[Theron *et al.* 2014a]
	Ethambutol	*embB,*	DNA from clinical specimen (sputum)	55 (95% CI 47-63)	78 (95% CI 69-85)	Not reported	Meta-analysis	[Cheng *et al.* 2014]
			Purified DNA from culture	64 (95% CI 60-67)	70 (95% CI 67-74)	Not reported	Meta-analysis	(Cheng *et al.* 2014)
AID TB Resistance (AID Diagnostika)	Rifampin	*rpoB*	DNA from clinical specimen (respiratory, 95% smear positive)	100 (95% CI 89.8-99.0)	100 (95% CI 77.1-100)	Not reported, "similar to MTBDR*plus*/ MDRTB*sl*"	Individual study	[Molina-Moya *et al.* 2015]
			DNA from clinical specimen (respiratory +	***	100 (95% CI 95.9-100)	<1 d	Individual study	[Ritter *et al.* 2014]

		non-respiratory, smear positive)					
		MGIT culture	100% (95% CI 29-100)	100% (95% CI 92-100)	<1 d	Individual study	[Ritter *et al.* 2014]
Isoniazid	*katG, inhA*	DNA from clinical specimen (respiratory, 95% smear positive)	97.8 (95% CI 87.0-99.9)	100 (95% CI 73.2-100)	Not reported, "similar to MTBDR*plus*/ MDRTB*sl*"	Individual study	[Molina-Moya *et al.* 2015]
		DNA from clinical specimen (respiratory + non-respiratory, smear positive)	***	100 (95% CI 95.9-100)	<1 d	Individual study	[Ritter *et al.* 2014]
		MGIT culture	100% (95% CI 29-100)	100% (95% CI 92-100)	<1 d	Individual study	[Ritter *et al.* 2014]**
Fluoroquinolones	*gyrA*	DNA from clinical specimen (respiratory, 95% smear positive)	33.3 (95% CI 6.0-75.9)	98.1 (95% CI 88.6-99.9)	Not reported, "similar to MTBDR*plus*/ MDRTB*sl*"	Individual study	[Molina-Moya *et al.* 2015]

		DNA from clinical specimen (respiratory + non-respiratory, smear positive)	No resistance	100 (95% CI 88-100)	<1 d	Individual study	[Ritter *et al.* 2014]
Ethambutol	*embB*	DNA from clinical specimen (respiratory, 95% smear positive)	60.0 (95% CI 42.2-75.6)	91.7 (95% CI 71.5-98.5)	Not reported, "similar to MTBDR*plus*/ MDRTB*sl*"	Individual study	[Molina-Moya *et al.* 2015]
		DNA from clinical specimen (respiratory + non-respiratory, smear positive)	100 (95% CI 3-100)	100 (95% CI 87.7-100)	<1 d	Individual study	[Ritter *et al.* 2014]
Aminoglycosides (kanamycin and capreomycin)	*rrs*	DNA from clinical specimen (respiratory, 95% smear positive)	100 (95% CI 77.1-100)	100 (95% CI 87.4-100)	Not reported, "similar to MTBDR*plus*/ MDRTB*sl*"	Individual study	[Molina-Moya *et al.* 2015]
		DNA from clinical specimen (respiratory + non-respiratory, smear positive)	-	100 (95% CI 89.7-100)	<1 d	Individual study	[Ritter *et al.* 2014]

Streptomycin	*RpsL, rrs*	DNA from clinical specimen (respiratory, 95% smear positive)	100 (95% CI 81.5-100)	96.6 (95% CI 80.4-99.8)	Not reported, "similar to MTBDR*plus*/ MDRTB*sl*"	Individual study	[Molina-Moya *et al.* 2015]
		DNA from clinical specimen (respiratory + non-respiratory, smear positive)	-	100 (95% CI 89.7-100)	<1 d	Individual study	[Ritter *et al.* 2014]

For meta-analyses: if available, ranges are shown in addition to pooled estimates, to indicate potential heterogeneity. All tests shown, with exception of GeneXpert, are line probe assays. Where no sensitivity is reported, no isolates were identified with resistance to the target drug. *Includes studies using MTBDR (first-generation). **Results not shown for second-line drugs, as only testing was only conducted on the 3 samples with resistance to first-line drugs. AID predicted 3/3 isolates to be susceptible to second-line, confirmed with phenotypic DST. ***Study did not report separate test results for positive RIF and INH resistance.

References

Altschul, S., Gish, W., Miller, W., Myers, E. and Lipman, D. (1990) Basic local alignment search tool. J Mol Biol 215: 403-410.

Alvarez, G., Van Dyk, D., Desjardins, M., Yasseen, A. III, Aaron, S., Cameron, D. *et al*. (2015) The feasibility, accuracy, and impact of Xpert MTB/RIF testing in a remote aboriginal community in Canada. Chest 148: 767-773.

Behr, M., Warren, S., Salamon, H., Hopewell, P., Ponce de Leon, A., Daley, C. *et al*. (1999) Transmission of *Mycobacterium tuberculosis* from patients smear-negative for acid-fast bacilli. Lancet 353: 444-449.

Boehme, C., Nabeta, P., Hillemann, D., Nicol, M., Shenai, S., Krapp, F. *et al*. (2010) Rapid molecular detection of tuberculosis and rifampin resistance. N Engl J Med 363: 1005-1015.

Brown, A., Bryant, J., Einer-Jensen, K., Holdstock, J., Houniet, D., Chan, J. *et al*. (2015) Rapid whole-genome sequencing of *Mycobacterium tuberculosis* isolates directly from clinical samples. J Clin Microbiol 53: 2230-2237.

Bryant, J., Schurch, A., van Deutekom, H., Harris, S., de Beer, J., de Jager, V. *et al*. (2013) Inferring patient to patient transmission of *Mycobacterium tuberculosis* from whole genome sequencing data. BMC Infect Dis 13: 110.

Carpentier, E., Drouillard, B., Dailloux, M., Moinard, D., Vallee, E., Dutilh, B. *et al*. (1995) Diagnosis of tuberculosis by Amplicor *Mycobacterium Tuberculosis* Test - a multicenter study. J Clin Microbiol 33: 3106-3110.

Casali, N., Nikolayevskyy, V., Balabanova, Y., Harris, S., Ignatyeva, O., Kontsevaya, I. *et al*. (2014) Evolution and transmission of drug-resistant tuberculosis in a Russian population. Nat Genet 46: 279-286.

Centers for Disease Control and Prevention (CDC). (2009) Updated guidelines for the use of Nucleic Acid Amplification Tests in the diagnosis of tuberculosis. MMWR Morb Mortal Wkly Rep 58: 7-10.

Cheng, S., Cui, Z., Li, Y. and Hu, Z. (2014) Diagnostic accuracy of a molecular drug susceptibility testing method for the antituberculosis drug ethambutol: a systematic review and meta-analysis. J Clin Microbiol 52: 2913-2924.

Chihota, V., Grant, A., Fielding, K., Ndibongo, B., van Zyl, A., Muirhead, D. *et al*. (2010) Liquid versus solid culture for tuberculosis: performance and cost in a resource-constrained setting. Int J Tuberc Lung Dis 14: 1024-1031.

Clinical and Laboratory Standards Institute. (2014) Nucleic acid sequencing methods in diagnostic laboratory medicine; Approved guideline - second edition. CLSI document MM09-A2. Clinical and Laboratory Standards Institute: Wayne, PA.

Cole, S., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D. *et al*. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 393: 537-544.

Coll, F., McNerney, R., Preston, M., Afonso Guerra- Assuncao, J., Warry, A., Hill-Cawthorne, G. *et al*. (2015) Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. Genome Med 7: 51.

Comas, I., Chakravartti, J., Small, P., Galagan, J., Niemann, S., Kremer, K. *et al*. (2010) Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. Nat Genet 42: 498-503.

Demers, A., Verver, S., Boulle, A., Warren, R., van Helden, P., Behr, M. *et al*. (2012) High yield of culture-based diagnosis in a TB-endemic setting. BMC Infect Dis 12: 218.

Denkinger, C., Schumacher, S., Boehme, C., Dendukuri, N., Pai, M. and Steingart, K. (2014) Xpert MTB/RIF Assay for the diagnosis of extrapulmonary tuberculosis: a systematic review and meta-analysis. Eur Respir J 44: 435-446.

Doughty, E., Sergeant, M., Adetifa, I., Antonio, M. and Pallen, M. (2014) Culture-independent detection and characterization of *Mycobacterium tuberculosis* and *M. Africanum* in sputum samples using shotgun metagenomics on a benchtop sequencer. Peer J 2(3): e585, DOI: 10.7717/peerj.585/supp-3.

Drobniewski, F., Nikolayevskyy, V., Maxeiner, H., Balabanova, Y., Casali, N., Kontsevaya, I. *et al*. (2013) Rapid diagnostics of tuberculosis and drug resistance in the industrialized world: clinical and public health benefits and barriers to implementation. BMC Med 11: 190.

Fadzilah, M., Peng Ng, K. and Fong Ngeow, Y. (2009) The manual MGIT system for the detection of *M. tuberculosis* in respiratory specimens: an experience in the University Malaya Medical Centre. Malay J Pathol 31: 93-97.

Farhat, M., Shapiro, B., Kieser, K., Sultana, R., Jacobson, K., Victor, T. *et al*. (2013) Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. Nat Genet 45: 1183-1189.

Feuerriegel, S., Schleusener, V., Beckert, P., Kohl, T., Miotto, P., Cirillo, D. *et al*. (2015) PhyResSE: a web tool delineating *Mycobacterium tuberculosis* antibiotic resistance and lineage from whole-genome sequencing data. J Clin Microbiol 53: 1908-1914.

Flandrois, J., Lina, G. and Dumitrescu, O. (2014) MUBII-TB-DB: a database of mutations associated with antibiotic resistance in *Mycobacterium tuberculosis*. BMC Bioinform 15: 107.

Gardy, J., Johnston, J., Ho Sui, S., Cook, V., Shah, L., Brodkin, E. *et al*. (2011) Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N Engl J Med 364: 730-739.

Guerra-Assuncao, J., Crampin, A. and Houben, R. (2015) Large-scale whole genome sequencing of *M. tuberculosis* provides insights into transmission in a high prevalence area. Elife 4: e05166.

Harris, S., Torok, M., Cartwright, E., Quail, M., Peacock, S. and Parkhill, J. (2013) Read and assembly metrics inconsequential for clinical utility of whole-genome sequencing in mapping outbreaks. Nat Biotechnol 31: 592-594.

Heather, J. and Chan, B. (2015) The sequence of sequencers: The history of sequencing DNA. Genomics DOI: 10.1016/j.ygeno.2015.11.003.

Hewlett, D., Horn, D. and Alfalla, C. (1995) Drug- resistant tuberculosis: Inconsistent results of pyrazinamide susceptibility testing. JAMA 273: 916-917.

Horne, D., Pinto, L., Arentz, M., Lin, S., Desmond, E., Flores, L. *et al*. (2013) Diagnostic accuracy and reproducibility of WHO-endorsed phenotypic drug susceptibility testing methods for first-line and second-line antituberculosis drugs. J Clin Microbiol 51: 393-401.

Ichiyama, S., Iinuma, Y., Yamori, S., Hasegawa, Y., Simokata, K. and Nakashima, N. (1997) Mycobacterium growth indicator tube testing in conjunction with the AccuProbe or the AMPLICOR-PCR assay for detecting and identifying mycobacteria from sputum samples. J Clin Microbiol 35: 2022-2025.

Jain, M., Fiddes, I., Miga, K., Olsen, H., Paten, B. and Akeson, M. (2015) Improved data analysis for the MinION Nanopore sequencer. Nat Meth 12: 351-356.

Jamieson, F., Guthrie, J., Neemuchwala, A., Lastovetska, O., Melano, R. and Mehaffy, C. (2014a) Profiling of *rpoB* mutations and MICs for rifampin and rifabutin in *Mycobacterium tuberculosis*. J Clin Microbiol 52: 2157-2162.

Jamieson, F., Teatero, S., Guthrie, J.L., Neemuchwala, A., Fittipaldi, N. and Mehaffy, C. (2014b) Whole- genome sequencing of the *Mycobacterium tuberculosis* Manila sublineage results in less

clustering and better resolution than Mycobacterial Interspersed Repetitive Unit Variable Number Tandem Repeat (MIRU-VNTR) typing and spoligotyping. J Clin Microbiol 52: 3795-3798.

Kato-Maeda, M., Ho, C., Passarelli, B., Banaei, N., Grinsdale, J., Flores, L. *et al*. (2013) Use of whole genome sequencing to determine the microevolution of *Mycobacterium tuberculosis* during an outbreak. PLoS One 8(3): e58235.

Köser CU, Bryant JM, Becq, J, Torok, ME, Ellington, MJ *et al*. (2013) Whole-genome sequencing for rapid susceptibility testing of M. tuberculosis. New Engl J Med 369(3):290–292.

Kwong, J., McCallum, N., Sintchenko, V. and Howden, B. (2015) Whole genome sequencing in clinical and public health microbiology. Pathology 47: 199-210.

Lander, E. and Waterman, M. (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics 2: 231-239.

Laver, T., Harrison, J., O'Neill, P., Moore, K., Farbos, A., Paszkiewicz, K. *et al*. (2015) Assessing the performance of the Oxford Nanopore technologies MinIon. Biomol Detect Quantif 3: 1-8.

Lee, R., Radomski, N., Proulx, J., Manry, J., McIntosh, F., Desjardins, F. *et al*. (2015) Reemergence and amplification of tuberculosis in the Canadian arctic. J Infect Dis 211: 1905-1914.

Ling, D., Zwerling, A. and Pai, M. (2008) GenoType MTBDR assays for the diagnosis of multidrug-resistant tuberculosis: a meta-analysis. Eur Respir J 32: 1165-1174.

Loman, N., Constantinidou, C., Chan, J., Halachev, M., Sergeant, M., Penn, C. *et al*. (2012a) High- throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. Nat Rev Microbiol 10: 599-606.

Loman, N., Misra, R., Dallman, T., Constantinidou, C., Gharbia, S., Wain, J. *et al*. (2012b) Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol 30: 434-439.

Long, S., Beres, S., Olsen, R. and Musser, J. (2014) Absence of patient-to-patient intrahospital transmission of *Staphylococcus aureus* as determined by whole-genome sequencing. mBio 5(5): e01692-e1714.

Maynard-Smith, L., Larke, N., Peters, J. and Lawn, S. (2014) Diagnostic accuracy of the Xpert MTB/RIF assay for extrapulmonary and pulmonary tuberculosis when testing non-respiratory samples: a systematic review. BMC Infect Dis 14: 709.

Migliori, G., Matteelli, A., Cirillo, D. and Pai, M. (2008) Diagnosis of multidrug-resistant tuberculosis and extensively drug-resistant tuberculosis: current standards and challenges. Can J Infect Dis Med Microbiol 19: 169-172.

Mikheyev, A. and Tin, M. (2014) A first look at the Oxford Nanopore MinION Sequencer. Mol Ecol Resour 14: 1097-1102.

Molina-Moya, B., Lacoma, A., Prat, C., Diaz, J., Dudnyk, A., Haba, L. *et al*. (2015) AID TB Resistance line probe assay for rapid detection of resistant *Mycobacterium tuberculosis* in clinical samples. J Infect 70: 400-408.

Morgan, M., Kalantri, S., Flores, L. and Pai, M. (2005) A commercial line probe assay for the rapid detection of rifampicin resistance in *Mycobacterium tuberculosis*: a systematic review and meta-analysis. BMC Infect Dis 5(1): 62.

Nebenzahl-Guimaraes, H., Jacobson, K., Farhat, M. and Murray, M. (2014) Systematic review of allelic exchange experiments aimed at identifying mutations that confer drug resistance in *Mycobacterium tuberculosis*. J Antimicrobial Chemother 69: 331-342.

Noor, K., Shephard, L. and Bastian, I. (2015) Molecular diagnostics for tuberculosis. Pathology 47: 250-256.

Ocheretina, O., Shen, L., Escuyer, V., Mabou, M., Royal-Mardi, G., Collins, S. *et al*. (2015) Whole genome sequencing investigation of a tuberculosis outbreak in Port-Au-Prince, Haiti caused by a strain with a 'Low-Level' *rpoB* mutation L511P - insights into a mechanism of resistance escalation. PLoS One 10(6): e0129207.

Olson, N., Lund, S., Colman, R., Foster, J., Sahl, J., Schupp, J. *et al*. (2015) Best practices for evaluating single nucleotide variant calling methods for microbial genomics. Front Genet 6: 235.

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M. *et al*. (2014) A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform 15: 256-278.

Park, M., Davis, A., Schluger, N., Cohen, H. and Rom, W. (1996) Outcome of MDR-TB patients, 1983-1993 - prolonged survival with appropriate therapy. Am J Respir Crit Care Med 153: 317-324.

Parrish, N. and Carroll, K. (2008) Importance of improved TB diagnostics in addressing the extensively drug-resistant TB crisis. Future Microbiol 3: 405-413.

Parrish, N. and Carroll, K. (2011) Role of the clinical mycobacteriology laboratory in diagnosis and management of tuberculosis in low-prevalence settings. J Clin Microbiol 49: 772-776.

Perkins, M. and Cunningham, J. (2007) Facing the crisis: improving the diagnosis of tuberculosis in the HIV era. J Infect Dis 196(Suppl. 1): S15-S27.

Price, J., Golubchik, T., Cole, K., Wilson, D., Crook, D., Thwaites, G. *et al*. (2014) Whole-genome sequencing shows that patient-to-patient transmission rarely accounts for acquisition of

Staphylococcus aureus in an intensive care unit. Clin Infect Dis 58: 609-618.

Public Health Agency of Canada. (2015) Tuberculosis: Drug resistance in Canada. 2013, Minister of Public Works and Government Services Canada: Ottawa (Canada).

Quick, J., Cumley, N., Wearn, C., Niebel, M., Constantinidou, C., Thomas, C. *et al*. (2014) Seeking the source of *Pseudomonas aeruginosa* infections in a recently opened hospital: an observational study using whole-genome sequencing. BMJ Open 4: e006278.

Quick, J., Ashton, P., Calus, S., Chatt, C., Gossain, S., Hawker, J. *et al.* (2015) Rapid draft sequencing and real-time Nanopore sequencing in a hospital outbreak of *Salmonella*. Genome Biol 16: 114.

Ritter, C., Lucke, K., Sirgel, F., Warren, R., van Helden, P., Bottger, E. *et al*. (2014) Evaluation of the AID TB Resistance line probe assay for rapid detection of genetic alterations associated with drug resistance in *Mycobacterium tuberculosis* strains. J Clin Microbiol 52: 940-946.

Roetzer, A., Diel, R., Kohl, T., Ruckert, C., Nubel, U., Blom, J. et al. (2013) Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiological study. PLoS Med 10(2): e1001387.

Sanchez-Padilla, E., Merker, M., Beckert, P., Jochims, F., Diamini, T., Kahn, P. *et al*. (2015) Detection of drug-resistant tuberculosis by Xpert MTB/RIF in Swaziland. N Engl J Med 372: 1181-1182.

Sandgren, A., Strong, M., Muthukrishnan, P., Weiner, B., Church, G. and Murray, M. (2009) Tuberculosis drug resistance mutation database. PLoS Med 6(2): e1000002.

Schurch, A., Kremer, K., Daviena, O., Kiers, A., Boeree, M., Siezen, R. *et al*. (2010) High-resolution typing by integration of genome sequencing data in a large tuberculosis cluster. J Clin Microbiol 48: 3403-3406.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O. and Huttenhower, C. (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Meth 9: 811-814.

SenGupta, D., Cummings, L., Hoogestraat, D., Butler-Wu, S., Shendure, J., Cookson, B. *et al*. (2014) Whole-genome sequencing for high-resolution investigation of methicillin-resistant *Staphylococcus aureus* epidemiology and genome plasticity. J Clin Microbiol 52: 2787-2796.

Snitkin, E., Zelazny, A., Thomas, P., Stock, F., NISC Comparative Sequencing Program Group, Henderson, D. *et al*. (2012) Tracking a hospital outbreak of carbapenem-resistant *Klebsiella pneumoniae* with whole-genome sequencing. Sci Transl Med 4(148): 148ra116.

Sohn, H., Aero, A., Menzies, D., Behr, M., Schwartzman, K., Alvarez, G. *et al*. (2014) Xpert MTB/RIF testing in a low tuberculosis incidence, high-resource setting: limitations in accuracy and clinical impact. Clin Infect Dis 58: 970-976.

Steiner, A., Stucki, D., Coscolla, M., Borell, S. and Gagneux, S. (2014) KvarQ: targeted and direct variant calling from fastq reads of bacterial genomes. BMC Genom 15: 881.

Steingart, K., Schiller, I., Horne, D., Pai, M., Boehme, C. and Dendukuri, N. (2014) Xpert MTB/RIF assay for pulmonary tuberculosis and rifampicin resistance in adults (review). Cochrane Libr 1: 1-168.

Steingart, K., Henry, M., Ng, V., Hopewell, P., Ramsay, A., Cunningham, J., Urbanczik, R. *et al*. (2006a) Fluorescence versus conventional sputum smear microscopy for tuberculosis: a systematic review. Lancet Infect Dis 6: 570-581.

Steingart, K., Ng, V., Henry, M., Hopewell, P., Ramsay, A., Cunningham, J., Urbanczik, R. *et al*. (2006b) Sputum processing methods to improve the sensitivity of smear microscopy for tuberculosis: a systematic review. Lancet Infect Dis 6: 664-674.

Stucki, D., Ballif, M., Bodmer, T., Coscolla, M., Maurer, A., Droz, S. *et al.* (2015) Tracking a tuberculosis outbreak over 21 years: strain-specific single- nucleotide polymorphism typing combined with targeted whole-genome sequencing. J Infect Dis 211: 1306-1316.

Theron, G., Peter, J., Richardson, M., Barnard, M., Donegan, S., Warren, R. *et al*. (2014a) The diagnostic accuracy of the GenoType MTBDRsl assay for the detection of resistance to second-line anti-tuberculosis drugs (review). Cochrane Libr 10: 1-123.

Theron, G., Zijenah, L., Chanda, D., Clowes, P., Rachow, A., Lesosky, M. *et al.* (2014b) Feasibility, accuracy, and clinical effect of point-of-care XpertMTB/RIF testing for tuberculosis in primary care settings in Africa: a multicentre, randomised, controlled trial. Lancet 383: 424-435.

Thorvaldsdottir, H., Robinson, J. and Mesirov, J. (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 14: 178-192.

Torok, M., Reuter, S., Bryant, J., Koser, C., Stinchcombe, S., Nazareth, B. *et al*. (2013) Rapid whole-genome sequencing for investigation of a suspected tuberculosis outbreak. J Clin Microbiol 51: 611-614.

Trauner, A., Borrell, S., Reither, K. and Gagneux, S. (2014) Evolution of drug resistance in tuberculosis: recent progress and implications for diagnosis and therapy. Drugs 74: 1063-1072.

Vuorinen, P., Miettinen, A., Vuento, R. and Hallstrom, O. (1995) Direct detection of *Mycobacterium tuberculosis* complex in respiratory specimens by Gen-Probe Amplified *Mycobacterium tuberculosis* Direct Test and Roche Amplicor *Mycobacterium Tuberculosis* Test. J Clin Microbiol 33: 1856-1859.

Walker, T., Ip, C., Harrell, R., Evans, J., Kapatai, G., Dedicoat, M. *et al*. (2013) Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational

study. Lancet Infect Dis 13: 137-146.

Walker, T., Kohl, T., Omar, S., Hedge, J., Del Ojo Elias, C., Bradley, P. *et al*. (2015) Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: a retrospective cohort study. Lancet Infect Dis 15: 1193-1202.

Witney, A., Gould, K., Arnold, A., Coleman, D., Delgado, R., Dhillon, J. *et al*. (2015) Clinical application of whole-genome sequencing to inform treatment for multidrug-resistant tuberculosis cases. J Clin Microbiol 53: 1473-1483.

Wood, D. and Salzberg, S. (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15(3): R46.

World Health Organization. (2015) Global Tuberculosis Report 2014, World Health Organization: Geneva.

Zetola, N., Shin, S., Tumedi, K., Moeti, K., Ncube, R., Nicol, M. *et al*. (2014) Mixed *Mycobacterium tuberculosis* complex infections and false-negative results for rifampin resistance by GeneXpert MTB/ RIF are associated with poor clinical outcomes. J Clin Microbiol 52: 2422-2429.

9.3 Additional unpublished analyses

Following this review, it is worth noting that a pilot study was published testing the use of WGS for clinical diagnostics (12). In this study, conducted at 8 different sites in the UK, Canada, Germany, Ireland and France, all newly positive MGIT samples were subjected to WGS using Illumina MiSeq. While sequencing was done on-site at these locations, resultant sequence data were uploaded to a server, with all bioinformatics done at a centralized location in the UK. WGS was used for speciation, detection of drug resistance and transmission. This approach was compared to conventional methods, which included smear (with or without GeneXpert), positive MGIT and then culture for diagnosis of *M. tuberculosis* complex (MTBC), followed by phenotypic drug susceptibility testing (DST), and then MIRU for epidemiologic analysis. For the purposes of this discussion (as in the manuscript), only diagnosis and prediction of drug resistance are considered, as these are required imminently by the treating physician.

Compared to conventional diagnostic methods (Hain Genotype MTBC/CM/AS), WGS had sensitivity of 95% (95% CI 91-98) and specificity of 98% (95% CI 95-100) for MTBC, including duplicate specimens. For resistance prediction, a list of resistance-connoting mutations was produced from Hain line probe assays and a review of the literature. Based on the presence or absence of these mutations in the genomes under investigation, 93% of genotypic results for first-line drugs agreed with phenotypic DST. However, WGS was unable to predict either resistance or sensitivity across 63 times across 15 samples due to insufficient depth of coverage. WGS also did not identify resistance-conferring mutations in 7 phenotypically resistant samples, 6 of which had 'unclassified' variants in the same genes as the *a priori* resistance-connoting mutations.

In this study, authors claimed a reduction in time to reporting of resistance results with WGS, compared to conventional DST. However, it was noted that this was a 'best-case scenario', wherein authors excluded the real delays they experienced due to batching of isolates for sequencing and the subsequent delays they experienced in uploading of these sequences to the server for centralized bioinformatics analysis. Without such adjustment, the median time from positive MGIT to DST report was 25 days for conventional DST (inter-quartile range, IQR 14-32) and 31 days from positive MGIT to MIRU report (IQR 21-44), compared to 31 days (IQR

21-60) for WGS with resistance and epidemiologic reporting. These results indicated that there was no observed reduction in time; rather there is a theoretical opportunity to achieve WGS-based predictions sooner than conventional DST. This suggests that until such time as specimens can be processed individually (such as using the Oxford Nanopore MinIon) and bioinformatics can be performed locally, conventional diagnostic approaches will continue to be faster.

CHAPTER 10. DISCUSSION AND CONCLUSIONS

The following chapter summarizes key findings from each Objective and discusses the public health implications of this work. I have also addressed the potential methodological considerations, and proposed directions for future research.

10.1 Discussion

The WHO has named Canada as a prime site for TB elimination based on its low overall incidence of this disease (63). However, Aboriginal-Canadians continue to experience rates of TB similar to those in developing nations, with the majority of cases thought to be due to ongoing transmission. This study was initiated in response to a major 'outbreak' in one such Aboriginal population, namely, the Inuit. As a previous molecular epidemiologic study in the Arctic had illustrated low strain diversity (2), and all isolates had the same DNA fingerprint based on classical typing methods, we employed a new technique, whole genome sequencing to increase resolution of this event. In conducting this study, key methodological issues were identified and explored, providing valuable insight into the application of this approach for epidemiology of TB, as well as other infectious diseases.

10.1.1 Summary of manuscripts and implications for public health

Manuscript I

The aim of this manuscript was to resolve transmission during the 'outbreak'. In contrast to the single group of transmission suggested by classical typing methods (MIRU and RFLP), WGS revealed three distinct clusters. Combining WGS with epidemiologic data then revealed at least 6 different subgroups ranging from a single instance of reactivation – inadvertently classified as part of the outbreak - to a point-source outbreak of 20. Small sample sizes within the subgroups precluded robust statistical analyses, however, by examining at this level, precise transmission events could be discerned. While one subgroup consisted of 5 teenagers who had socialized at local 'gathering houses', most subgroups were comprised of a mixture of household and social contacts. While some cases among the latter resided in such gathering houses, others did not, suggesting these houses were not the only sites of transmission as initially hypothesized by public health. By examining the contact investigation data in light of WGS, it was found that

once investigation moved beyond the household, the probability of detecting transmission was no better than chance. Finally, by comparing 'outbreak' isolates with historical isolates from this community, WGS was able to identify a case of recurrent TB due to reinfection that would otherwise have been classified as relapse.

Manuscript II

During the 'outbreak', there was substantial concern about whether transmission was occurring between villages – which would warrant rescheduling or cancelling of cross-community cultural events – and whether a new, hyper-virulent strain was responsible for the sudden increase of TB in this region. To address these concerns, this manuscript extends the previous investigation to other villages of Nunavik, applying WGS to cases diagnosed between 1990-2013. Examining pairwise SNP distributions within and between villages suggested transmission was mainly within villages, while evolutionary studies revealed the predominant strain (affecting 153/163 cases with WGS) has been circulating in Nunavik since the early 20th century. Surprisingly, this strain has thrived in Nunavik despite the accumulation of potentially deleterious nonsynonymous SNPs and deletions, a finding in direct contrast to other studies, which have suggested that epidemiologic success is a consequence of strain characteristics (e.g., (182)).

Manuscript III

During the analysis of the 'outbreak', it was observed that numerous cases had multiple contacts with different genotypes (as identified by the subgroups described in Manuscript I) as well as different potential sources. As two case-control studies (10, 11) were not able to sufficiently explain the high attack rate in this community, it was hypothesized that multiple exposures may have influenced progression. Adjusting for housing occupancy, increased total exposures (i.e., exposures to any confirmed case) were associated with progression to disease among those with recent infection. This suggests that the degree to which one is exposed is not only associated with the risk of initial infection with TB, but also the risk of progression from that infection to disease.

Manuscript IV

Given the increasing use of WGS in epidemiologic studies of TB, as well as other pathogens, this manuscript aimed to explore the potential influence of different reference genomes on phylogenetic trees and subsequent epidemiologic inferences. Using 7 different reference genomes of increasing divergence from the Inuit isolates (Manuscripts I-III), it was shown that clusters could still be resolved even using *Mycobacterium bovis* as a reference, with a single isolate changing clusters when using the even more distantly-related *M. canettii* (with 96% average nucleotide identity versus a *M. tuberculosis* lineage 4 reference) as a reference. Capitalizing on this unique low-diversity dataset, such results can easily be extrapolated to other settings, where *M. tuberculosis* strain diversity is typically much higher.

Manuscript V

As the role of WGS in tuberculosis epidemiology (as well as other infectious diseases) has become solidified, its use as a diagnostic tool has recently been proposed. In response to this, we were requested to write a narrative review on this subject. While 2 studies demonstrated it is possible to obtain and sequence *M. tuberculosis* from raw clinical specimens, one was not able to obtain sufficient depth of coverage for other analyses, while the other relied on a costly and labour-intensive protocol for DNA enrichment. Therefore, at present, WGS is performed using DNA extracted from culture. This precludes its utility in clinical diagnosis in high-resource settings, wherein smear microscopy, a DNA amplification test and DNA probe would have already provided a diagnosis and the patient has typically already started treatment. Several studies have examined WGS for prediction of drug resistance; however, accuracy is limited by current knowledge of drug-resistance mutations and would result in substantial false positive results in settings such as Canada with low levels of drug resistance. Unfortunately, these are the contexts wherein WGS is presently most feasible. We concluded that at a minimum, WGS should be utilized as an adjunct to phenotypic drug susceptibility testing and not as a replacement test – a view supported by a TBNET/RESIST-TB consensus statement released in January 2016 (183).

10.1.2 Implications for TB control in Nunavik

Based on the findings of Manuscript I, several changes have been made in the public health

management of TB in Nunavik. Firstly, the low positive predictive value of contact investigation beyond the household and high staffing demands such investigation requires has prompted the NRBHSS to instead perform community-wide chest x-ray screening during outbreaks to detect prevalent cases. Such a screening was conducted in another Arctic community earlier this year. Secondly, given the demonstrated utility of WGS in differentiating closely-related strains of TB, and discriminating reactivation from reinfection, the NRBHSS has asked that we performed prospective WGS of TB cases from this community. During this follow-up, we have identified other cases of reinfection and an instance of reactivation of a strain not seen in the community since 2004, with the latter reinforcing the importance of early and complete LTBI prophylaxis in this context.

By revealing TB transmission is predominantly within villages in Nunavik (Manuscript II), our work suggests that even when during a TB epidemic, community quarantine is not required, nor is cancellation of cross-community events. As some villages have had no cases of TB in over 23 years, this also suggests that interventions can be village-specific, rather than applying the same TB control measures to all communities. For example, BCG vaccination was reinstituted in the 'outbreak' village as well as another community with elevated rates of TB. Furthermore, our work provides evidence against the introduction of a new, hyper-virulent strain in this region, suggesting clinical TB care in Nunavik can continue as previous with emphasis, as always, on adherence to both active TB treatment and LTBI prophylaxis.

The finding that increased exposure is not only associated with infection, but progression from such infection to disease (Manuscript III) reinforces the importance of LTBI prophylaxis in this context. While adherence based on monthly pill counts is high (generally >80%), options such as direct-observed therapy were made available to patients during the 'outbreak' and may continue to be useful in helping assure high adherence. Among those who decline LTBI prophylaxis, this finding suggests a need for increased clinical monitoring during the years immediately following infection, to ensure incident cases are detected early and secondary transmission is reduced.

10.1.3 Implications for WGS-based studies of TB and other infectious diseases

During the years over which this thesis work was conducted, a number of studies have examined

WGS for resolving TB transmission (e.g., (101, 104, 105, 107, 108, 184, 185)). Many have relied on a threshold approach, with a precise number of SNPs considered to support or refute transmission (101, 105, 109, 115, 184)To our knowledge, our study is the first to illustrate the importance of local strain diversity in the application and interpretation of such SNP thresholds. Furthermore, our study has highlighted the continued importance of epidemiologic data in addition to WGS data for resolving transmission. While WGS may be sufficient to rule out transmission – given sufficient genetic difference between pairs of isolates, and with consideration of local strain diversity – absent epidemiologic data, it cannot rule it in. In the Northern 'outbreak', which occurred over the course of a single year, many isolates had very few SNPs between them. WGS alone identified three clusters of transmission, confirming the findings of (3, 103, 104) that it provides greater resolution of transmission compared to classical genotyping methods. However, to delineate transmission networks further within these clusters and discriminate subgroups, epidemiologic data was essential.

As mentioned, a key issue with WGS that bioinformatics pipelines are not currently standardized, with many of the analytic choices lacking formal validation. Numerous WGS studies, using isolates of various lineages of *M. tuberculosis*, have all aligned reads to the genome of the H37Rv strain, which belongs to lineage 4. Our study has validated this decision. In addition, as all clustering was lost with *M. kansasii,* which still offered genomic coverage of 35%, this suggests restricting analyses to a small fraction of a genome can lead to substantial bias in epidemiologic inferences of transmission. This has significance for other pathogens as well as *M. tuberculosis,* as approaches such as MLST - which rely on only a small subset of SNPs - are used extensively for surveillance.

10.1.4 Methodological considerations

A key limitation in all current WGS studies of tuberculosis is the detection of within-host strain diversity, either as a result of micro-evolution or mixed infection. As WGS was conducted using bacterial isolates obtained from cases at diagnosis, it is possible that within-host mutation could have occurred prior to diagnosis. Thus, a case might have initially transmitted an 'older' variant of the bacteria to some secondary cases, and over time, developed a mutation and transmitted this 'new' mutated bacteria to others. A case might also have been infected by >1 source,

resulting in two populations of bacteria within the lung.

Detection of such diversity may also be influenced by culture. All WGS was performed on DNA extracted from cultured *M. tuberculosis.* Martin *et al.* (186) experimentally combined 7 pairs of *M. tuberculosis* strains at different concentrations (90/10; 10/90; 95/5; 5/95 and 50/50) and subjected to these mixtures to MIRU both before and after 2-3 weeks of culture. In 4/7 pairs, culture resulted in significant changes in strain diversity, with mixed infection no longer detectable in 1 to all 5 of the mixtures. Given that WGS offers higher discrimination than MIRU, it seems plausible that mixed infections would be detectable at much lower thresholds. However, while a protocol for sequencing *M. tuberculosis* directly from the clinical sample has only recently been developed (187), no studies as yet have evaluated this. Furthermore, in our study, all colonies of *M. tuberculosis* visible on the plate were sequenced for each patient (i.e., a clean sweep was performed). This approach may result in differences in strain detection compared to sequencing individual colonies (152, 188). Additionally, as only one culture was available per patient, this may not fully reflect the diversity present in the native lung (189, 190) or that which was transmitted forward.

As precise transmission networks were investigated in the 'outbreak' village, minimizing the risk of missing such within-host heterogeneity was essential. To investigate this, we utilized a multi-faceted approach in agreement with recent recommendations (191), including manually inspecting the loci of cluster-defining SNPs for all isolates from this village. The maximum variation seen in any of our 'outbreak' isolates was 9% of reads containing a minority allele at a cluster-defining SNP. Deep sequencing of this isolate, as well as others selected based on detection of low-frequency variants and epidemiologic significance, revealed even lower proportions of minority variants. The percentage of alternative reads required to classify a locus as 'mixed' varies widely by study. Black *et al.* suggested that at least 30% of reads should have an alternative allele to be called mixed infection (or within-host evolution), yielding a true positive rate of 79.5% and false positive rate of 14.3% (188). While Guerra-Assuncao *et al.* used the same (123), others have used lower thresholds of 5 (117) or 10% (184). Therefore, while results suggest that mixed infection with strains from different clusters/subgroups or micro-evolution with strains representing >1 subgroup was highly unlikely, as the threshold for

identifying this remains unclear, this remains a possibility. Finally, because of the extremely low diversity within each subgroup (often 0 SNPs between some isolates), we were unable to assess for mixed infection from >1 source within these groups.

As WGS – by definition – requires a positive culture, another potential limitation is the requisite exclusion of clinical cases. Clinical cases are those diagnosed on the basis of chest radiograph and symptoms. During the 'outbreak', there were 19 clinical cases, with 42% under the age of 5 and 53% under the age of 10. While some of these may be misdiagnosed, others – particularly the pediatric patients – may be unable to produce sputum or have paucibacillary disease (6). These cases are generally not contagious and thus more likely to represent secondary cases than sources of transmission. However, as the source of transmission for pediatric cases is thought to be a close family contact (6), in most instances, probable source cases were still identified by public health. These clinical cases were not included in our analyses in Objective 3 because a) we aimed to evaluate genotypic exposures between all cases and contacts, b) some of these are unlikely to represent 'true' cases and c) the remainder likely represent a subpopulation with very different risk factors for progression compared to the study sample.

Another potential issue relates to sample size. Given that studies investigated transmission in a specific geographical area, we were limited by the absolute number of cases. However, in molecular/genomic epidemiology, the sampling fraction is more critical than absolute numbers to evaluate transmission. Time of sampling is also important, as too short a sampling period means source or secondary cases outside the sampling frame may be missed. As we were able to sequence 95% of all isolates from cases over 22 years in the 'outbreak' village, and over 90% of isolates across Nunavik from 2001-2013, with an additional sample of 26 isolates from 1990-2000, we expect sampling bias to be negligible.

All clinical and epidemiologic data used in this work were collected as part of a public health response, rather than for research purposes. As such, information regarding several epidemiologic risk factors was not collected consistently, limiting the variables that could be evaluated as potential risk factors for progression to TB disease in Objective 3. However, as there is little in- or out-migration from these communities, with life-long medical records

available, comprehensive demographic information and medical history were available and used to support the molecular epidemiologic analyses.

Measurement error in terms of the total (and consequently, genotypic) exposures used in Objective 3 is also possible. All individuals with active TB provided a list of their contacts at time of diagnosis. However, the precise details of this exposure were obtained from the interviews, wherein many of the contacts could not recall when they last had seen the recently-diagnosed case for whom they were being investigated. Additionally, the precise frequency or duration of this contact was not always available. However, as interviews with these contacts were generally conducted prior to their own diagnosis with either infection or active TB, knowledge of their own disease status should not have influenced reporting. Therefore, such error is likely to be non-differential between cases and controls. Additionally, as of May 2012, public health suspected transmission was occurring in local 'gathering houses' and announcements were made to this effect to the community. It is therefore possible that reporting attendance at such venues was subsequently influenced by social desirability bias. Time-stratified analyses, however, did not identify substantial differences in reported exposures.

10.1.5 Future directions

Since these studies, TB rates have continued to be elevated in Nunavik. While GeneXpert has been implemented in the two Northern hospitals to reduce diagnostic delay, monitoring for transmission outside of declared outbreaks continues to be passive. Given that one patient was symptomatic for 4 months prior to presenting to clinic, transmitting to at least 15 others in the 'outbreak' village (Manuscript I), this suggests patient delay (as opposed to provider or laboratory delay) may be a more important contributor to ongoing transmission. As such, active case finding in high-incidence villages such as that implemented in Alaska in the 1970s (68) may be warranted to halt transmission in this context. In addition, mixed methods studies examining patient delay and factors contributing to this may provide useful insights into the TB dilemma confronting the North. Given limited in- and out-migration in villages of Nunavik, complete case ascertainment and follow-up is possible. This presents a unique opportunity to monitor the success of public health interventions, using WGS to delineate recent transmission from reactivation. Ultimately, interventions in Nunavik may guide efforts to eliminate TB in other

low-incidence countries.

As WGS has become the gold standard for tracking infectious disease transmission, validation of bioinformatics pipelines used is essential. While one step has been addressed in this manuscript, there are many additional opportunities for investigation. For example, to date, no studies have examined the test parameters of various SNP calling algorithms using bacterial genomes. While previous studies have done so with human genetic data, there are substantial differences in depth of coverage, targets of sequencing (WGS versus whole exome sequencing) and even the goals of sequencing – all of which may influence how well these algorithms perform. Further study is also warranted into accurate detection of within-host heterogeneity, as this can not only impact estimates of transmission, but potentially individual patient outcomes as well (192). By experimentally mixing strains, as has been done in (186), the minimum depth of coverage needed to detect varying degrees of heterogeneity could be evaluated. Differences between transmitted strains and culture could also be assessed, as methods improve for sequencing *M. tuberculosis* DNA from raw clinical samples (e.g., directly from sputum).

Validation and standardization of such pipelines is also critical as public health agencies begin to implement WGS for routine surveillance. As these methods continue to require implementation via command line, specialized technical expertise is needed, with the demand vastly exceeding the trained personnel. Tools must be developed to simplify data analysis if WGS is to be conducted at local, rather than regional facilities. Furthermore, interpretation of WGS data currently requires an understanding of genomics, which are unfamiliar to most public health staff and clinicians. Therefore, approaches for communicating relevant results in an uncomplicated manner are also needed.

10.2 Overall conclusions

This project aimed to apply a newer method, WGS, to understanding TB transmission among the Inuit population of Nunavik. Where classical studies revealed limited strain diversity and could not accurately differentiate reactivation from recent transmission, WGS provided in-depth resolution of these events. This has increased our understanding of TB transmission in the North and informed public health interventions in this region. By examining methodological

considerations in the application of WGS, with implications not only for TB but other infectious diseases as well.

The high analytic power of WGS and declining cost suggest that this is the future of molecular epidemiology. Consequently, WGS is quickly moving from a research method to a tool for public health surveillance. However, this approach is currently limited to high-resource, low-incidence settings, where the majority of TB is thought to be due to reactivation rather than recent transmission. It is clear that, for resolving transmission, WGS would provide the greatest benefit in low-resource, high-incidence settings. Unfortunately, the current cost, resources and requisite technical and substantive expertise needed to perform WGS suggest that routine use of this method in such settings will not occur anytime in the near future. However, the targeted use of WGS in representative settings may be particularly valuable in understanding the drivers of TB transmission, and more importantly, transmission of drug-resistant TB, in low-resource, high-incidence countries.

REFERENCES

1. Cave MD, Eisenach KD, McDermott PF, Bates JH, Crawford JT. (1991) IS6110: conservation of sequence in the *Mycobacterium tuberculosis* complex and its utilization in DNA fingerprinting. *Mol Cell Probes* 5(1):73–80.

2. Nguyen D, *et al*. (2003) Tuberculosis in the Inuit community of Quebec, Canada. *Am J Respir Crit Care Med* 168(11):1353–1357.

3. Gardy JL, *et al*. (2011) Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *New Engl J Med* 364(8):730–739.

4. Grzybowski S, Barnett GD, Styblo K. (1975) Contacts of cases of active pulmonary tuberculosis. *Bull Int Union Tubercul Lung Dis* 50(1):90–106.

5. Ferguson RG. (1946) BCG vaccination in hospitals and sanatoria of Saskatchewan; a study carried out by the National Research Council of Canada. *Can J Public Health* 37(11):435–451.

6. Public Health Agency of Canada. (2014) *Canadian Tuberculosis Standards 7th Edition*.

7. Mack U, *et al*. (2009) LTBI: latent tuberculosis infection or lasting immune responses to M. tuberculosis? A TBNET consensus statement. *Eur Respir J* 33(5):956–973.

8. Sloot R, Schim van der Loeff MF, Kouw PM, Borgdorff MW. (2014) Risk of tuberculosis after recent exposure. A 10-Year Follow-up Study of Contacts in Amsterdam. *Amer J Respir Crit Care Med* 190(9):1044–1052.

9. Downes .J (1935) A study of the risk of attack among contacts in tuberculous families in a rural area. *Am J Epi* 22(3):731–742.

10. Fox GJ, *et al.* (2015) Inadequate Diet is associated with acquiring *Mycobacterium tuberculosis* infection in an Inuit community: a case-control study. *Annals ATS*:150622133645008.

11. Ahmed Khan F, *et al*. (2015) Housing characteristics as determinants of tuberculosis in an Inuit community: a case-control study. Presented at: The Union Against Tuberculosis and Lung Disease - North America Region (Vancouver).

12. Pankhurst LJ, *et al.* (2015) Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: a prospective study. *Lancet Respir Med* 4(1):49–58.

13. Sepkowitz KA (1996) How contagious is tuberculosis? *Clin Infect Dis* 23(5):954–962.

14. Mangura BTB, Napolitano ECE, Passannante MRM, McDonald RJR, Reichman LBL. (1998) *Mycobacterium tuberculosis* miniepidemic in a church gospel choir. *Chest* 113(1):234–237.

15. Ahmad S. (2011) Pathogenesis, immunology, and diagnosis of latent *Mycobacterium tuberculosis* infection. *Clin Dev Immunol* 2011(7):1–17.

16. Divangahi M. (2013) *The New Paradigm of Immunity to Tuberculosis* (Springer Science & Business Media).

17. Bustamante J, Boisson-Dupuis S, Abel L, Casanova J-L. (2014) Mendelian susceptibility to mycobacterial disease: Genetic, immunological, and clinical features of inborn errors of IFN-γ immunity. *Seminars in Immunology* 26(6):454–470.

18. Wu L. (2015) Screening toll-like receptor markers to predict latent tuberculosis infection and subsequent tuberculosis disease in a Chinese population. *BMC Med* (16):19.

19. Lin H-H, Ezzati M, Murray M. (2007) Tobacco smoke, indoor air pollution and tuberculosis: a systematic review and meta-analysis. *PLoS Med* 4(1):e20.

20. Sopori M (2002) Effects of cigarette smoke on the immune system. *Nat Rev Immunol* 2(5):372–377.

21. Keane J. (2016) Effects of cigarette smoke on the macrophage-pathogen interaction. *Keystone Symposia: Tuberculosis co-morbidities and immunopathogenesis*.

22. O'Leary SM, *et al.* (2014) Cigarette Smoking Impairs Human Pulmonary Immunity to Mycobacterium tuberculosis. *Amer J Respir Crit Care Med* 190(12):1430–1436.

23. Happel KI (2005) Alcohol, Immunosuppression, and the Lung. *Ann Am Thoracic Soc* 2(5):428–432.

24. Rehm J, *et al*. (2009) The association between alcohol use, alcohol use disorders and tuberculosis (TB). A systematic review. *BMC Public Health* 9(1):450.

25. Parry C, Rehm J, Poznyak V, Room R. (2009) Alcohol and infectious diseases: an overlooked causal linkage? *Addiction* 104(3):331–332.

26. Rieder HL. (1999) *Epidemiologic basis of tuberculosis control* (IUATAL).

27. Fox GJ, Barry SE, Britton WJ, Marks GB. (2012) Contact investigation for tuberculosis: a systematic review and meta-analysis. *Eur Respir J* 41(1):140–156.

28. Selwyn PA, *et al.* (1989) A prospective study of the risk of tuberculosis among intravenous drug users with human immunodeficiency virus infection. *New Engl J Med* 320(9):545–550.

29. Kwan CK, Ernst JD. (2011) HIV and tuberculosis: a deadly human syndemic. *Clin Microbiol Rev* 24(2):351–376.

30. Jeon CY, Murray MB. (2008) Diabetes mellitus increases the risk of active tuberculosis: a systematic review of 13 observational studies. *PLoS Med*.

31. Al-Efraij K, *et al.* (2015) Risk of active tuberculosis in chronic kidney disease: a systematic review and meta-analysis. *Int J Tuberc Lung Dis* 19(12):1493–1499.

32. Ganmaa D, *et al*. (2012) Vitamin D, tuberculin skin test conversion, and latent tuberculosis in Mongolian school-age children: a randomized, double-blind, placebo-controlled feasibility trial. *Am J Clin Nutr* 96(2):391–396.

33. Arnedo-Pena A, *et al.* (2011) Latent tuberculosis infection, tuberculin skin test and vitamin D status in contacts of tuberculosis patients: a cross-sectional and case-control study. *BMC Infect Dis* 11(1):349.

34. Arnedo-Pena A, *et al.* (2015) Vitamin D status and incidence of tuberculosis infection conversion in contacts of pulmonary tuberculosis patients: a prospective cohort study. *Epidemiol Infect* 143(8):1731–1741.

35. Marais BJ, Gie RP, Schaaf HS. (2004) The clinical epidemiology of childhood pulmonary tuberculosis: a critical review of literature from the pre-chemotherapy era. *Int J Tubercul Lung Dis* 8(3):287-285.

36. Weiskopf D, Weinberger B, Grubeck-Loebenstein B. (2009) The aging of the immune system. *Transpl Int* 22(11):1041–1050.

37. Steingart KR, *et al*. (2006) Fluorescence versus conventional sputum smear microscopy for tuberculosis: a systematic review. *Lancet Infect Dis* 6(9):570–581.

38. Yeager H, Lacy J, Smith LR, LeMaistre CA. (1967) Quantitative studies of mycobacterial populations in sputum and saliva. *Am Rev Respir Dis* 95(6):998–1004.

39. Styblo K. (1978) [Current status of the problem. I. Epidemiology of tuberculosis]. *Bull Intl Union Tubercul Lung Dis* 53(3):153–166.

40. Behr MA, *et al.* (1999) Transmission of *Mycobacterium tuberculosis* from patients smear-negative for acid-fast bacilli. *Lancet* 353(9151):444–449.

41. Yoder MA, Lamichhane G, Bishai WR. (2004) Cavitary pulmonary tuberculosis: the Holy Grail of disease transmission. *Cur Sci* 86(1):74–81.

42. Turner RD, Bothamley GH. (2015) Cough and the transmission of tuberculosis. *J Infect Dis* 211(9):1367–1372.

43. Loudon RG, Spohn SK. (1969) Cough frequency and infectivity in patients with

pulmonary tuberculosis. *Am Rev Respir Dis* 99(1):109–111.

44. Fennelly KP, *et al.* (2012) Variability of Infectious Aerosols Produced during Coughing by Patients with Pulmonary Tuberculosis. *Am J Respir Crit Care Med* 186(5):450–457.

45. Curtis, AB, *et al*. (1999) Extensive transmission of *Mycobacterium tuberculosis* from a child. *New Engl J Med* 341(20):1491–1495.

46. Nardell EA, Keegan J, Cheney SA, Etkind SC. (1991) Theoretical limits of protection achievable by building ventilation. *Am Rev Respir Dis* 144:302–306.

47. Urrego J, *et al*. (2015) The Impact of Ventilation and Early Diagnosis on Tuberculosis Transmission in Brazilian Prisons. *Am J Trop Med Hyg* 93(4):739–746.

48. Menzies D, Fanning A, Yuan L, FitzGerald JM. (2000) Hospital ventilation and risk for tuberculous infection in Canadian health care workers. Canadian Collaborative Group in Nosocomial Transmission of TB. *Ann Intern Med* 133(10):779–789.

49. Houk VN, Baker JH, Sorensen K, Kent DC. (1968) The epidemiology of tuberculosis infection in a closed environment. *Arch Environ Health* 16(1):26–35.

50. Gagneux S, *et al*. (2006) Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *Proc Natl Acad Sci USA* 103(8):2869–2873.

51. Firdessa R, *et al*. (2013) Mycobacterial lineages causing pulmonary and extrapulmonary tuberculosis, Ethiopia. *Emerging Infect Dis* 19(3):460–463.

52. López B, *et al.* (2003) A marked difference in pathogenesis and immune response induced by different *Mycobacterium tuberculos*is genotypes. *Clin Exp Immunol* 133(1):30–37.

53. Ford J, *et al.* (2013) *Mycobacterium tuberculosis* mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nat Genet* 45(7):784–790.

54. Cowley D, *et al*. (2008) Recent and rapid emergence of W-Beijing strains of *Mycobacterium tuberculosis* in Cape Town, South Africa. *Clin Infect Dis* 47(10):1252–1259.

55. Hanekom M, *et al.* (2007) A recently evolved sublineage of the *Mycobacterium tuberculosis* Beijing strain family is associated with an increased ability to spread and cause disease. *J Clin Micro* 45(5):1483–1490.

56. Wada T, *et al*. (2009) High transmissibility of the modern Beijing *Mycobacterium tuberculosis* in homeless patients of Japan . *Tuberculosis* 89(4):252–255.

57. Albanna AS, *et al.* (2011) Reduced transmissibility of East African Indian strains of *Mycobacterium tuberculosis*. *PLoS ONE* 6(9):e25075.

58. de Jong BC, *et al*. (2008) Progression to active tuberculosis, but Not transmission, varies by *Mycobacterium tuberculosis* lineage in The Gambia. *J Infect Dis* 198(7):1037–1043.

59. Comstock GW. (1999) How much isoniazid is needed for prevention of tuberculosis among immunocompetent adults? *Int J Tubercul Lung Dis* 3(10):847–850.

60. World Health Organization. (2015) *Global Tuberculosis Report 2014* (World Health Publications).

61. Gandhi NR, *et al*. (2010) Multidrug-resistant and extensively drug-resistant tuberculosis: a threat to global control of tuberculosis. *Lancet* 375(9728):1830–1843.

62. Kendall EA, Fofana MO, Dowdy DW (2015) Burden of transmitted multidrug resistance in epidemics of tuberculosis: a transmission modelling analysis. *Lancet Respir Med* 3(12):963–972.

63. World Health Organization. (2014) Framework towards ttuberculosis elimination in low-incidence countries.

64. Public Health Agency of Canada. (2016) Tuberculosis in Canada 2014 Pre-release. 1–20.

65. Kunimoto D, *et al*. (2004) Transmission characteristics of tuberculosis in the foreign-born and the Canadian-born populations of Alberta, Canada. *Int J Tuberc Lung Dis* 8(10):1213–1220.

66. Statistics Canada. (2013). Aboriginal Peoples in Canada: First Nations People, Métis and Inuit. *National Household Survey, 2011*:1–23. Catalogue no. 99-011-X2011001

67. Grygier PS. (1994) *A long way from home: the tuberculosis epidemic among the Inuit.* (McGill/Queen's University Press, Montreal).

68. Grzybowski S, Styblo K, Dorken E. (1976) Tuberculosis in Eskimos. *Tubercle* 57(4):S1–S58.

69. Wherrett GJ (1969) A study of tuberculosis in the Eastern Arctic. *Can J Public Health* 60(1):7–14.

70. Centers for Disease Control (1995) Essential Components of a Tuberculosis Prevention and Control Program: Recommendations of the Advisory Council for the Elimination of Tuberculosis. *CDC MMWR Morb Mort Wkly Rep* 44(RR-11).

71. Reichler MR, *et al*. (2002) Evaluation of investigations conducted to detect and

prevent transmission of tuberculosis. *JAMA* 287(8):991-995.

72. Veen J. (1992) Microepidemics of tuberculosis: the stone-in-the-pond principle. *Tuber Lung Dis* 73(2):73–76.

73. Faccini M, *et al*. (2015) Tuberculosis-related stigma leading to an incomplete contact investigation in a low-incidence country. *Epidemiol Infect* 143:2841–2848.

74. Diel R, Meywald-Walter K, Gottschalk R, Rusch-Gerdes S, Niemann S (2004) Ongoing outbreak of tuberculosis in a low-incidence community: a molecular-epidemiological evaluation. *Int J Tubercul Lung Dis* 8(7):855–861.

75. Asghar RJ, *et al*. (2009) Limited utility of name-based tuberculosis contact investigations among persons using illicit drugs: results of an outbreak investigation. *J Urban Health* 86(5):776–780.

76. Behr MA, *et al*. (1998) Predictive value of contact investigation for identifying recent transmission of *Mycobacterium tuberculosis*. *Am J Respir Crit Care Med* 158(2):465–469.

77. Hermans PW, *et al*. (1990) Insertion element IS986 from *Mycobacterium tuberculosis*: a useful tool for diagnosis and epidemiology of tuberculosis. *J Clin Micro* 28(9):2051–2058.

78. van Soolingen D, De Haas P, Hermans P, van Embden JD (1994) DNA Fingerprinting of mycobacterium tuberculosis. *Methods Enzymol* 235:196–205.

79. Daley CL, *et al.* (1992) An outbreak of tuberculosis with accelerated progression among persons infected with the human immunodeficiency virus. An analysis using restriction-fragment-length polymorphisms. *New Engl J Med* 326(4):231–235.

80. de Boer AS, *et al.* (1999) Analysis of rate of change of IS*6110* RFLP patterns of *Mycobacterium tuberculosis* based on serial patient isolates. *J Infect Dis* 180(4):1238–1244.

81. Rosenberg NA, Tsolaki AG, Tanaka MM (2003) Estimating change rates of genetic markers using serial samples: applications to the transposon IS6110 in *Mycobacterium tuberculosis*. *Theor Popul Biol* 63(4):347–363.

82. Barnes PF, Cave MD. (2003) Molecular epidemiology of tuberculosis. *New Engl J Med* 349(12):1149–1156.

83. Roetzer A, *et al*. (2011) Evaluation of Mycobacterium tuberculosis typing methods in a 4-Year study in Schleswig-Holstein, Northern Germany. *J clin microbio* 49(12):4173–4178.

84. Jonsson J, *et al*. (2014) Comparison between RFLP and MIRU-VNTR Genotyping of *Mycobacterium tuberculosis* strains isolated in Stockholm 2009 to 2011. *PLoS*

ONE 9(4):e95159.

85. Warren RM, *et al.* (2002) Microevolution of the Direct Repeat region of *Mycobacterium tuberculosis*: implications for interpretation of spoligotyping data. *J clin micro* 40(12):4457–4465.

86. Sepkowitz KA, *et al*. (1995) Tuberculosis among urban health care workers: a study using restriction fragment length polymorphism typing. *Clin Infect Dis* 21(5):1098–1101.

87. Braden CR, *et al.* (1997) Retrospective detection of laboratory cross-contamination of *Mycobacterium tuberculosis* cultures with use of DNA fingerprint analysis. *Clin Infect Dis* 24(1):35–40.

88. Cook VJ, Stark G, Roscoe DL, Kwong A, Elwood RK (2014) Investigation of suspected laboratory cross-contamination: interpretation of single smear-negative, positive cultures for *Mycobacterium tuberculosis*. *CMI* 12(10):1042–1045.

89. Jasmer RM, *et al*. (1999) A molecular epidemiologic analysis of tuberculosis trends in San Francisco, 1991-1997. *Ann Intern Med* 130(12):971–978.

90. Iñigo J, Arce A, Palenque E, García de Viedma D, Chaves F (2008) Decreased tuberculosis incidence and declining clustered case rates, Madrid. *Emerging Infect Dis* 14(10):1641–1643.

91. Borgdorff MW, *et al*. (2010) Progress towards tuberculosis elimination: secular trend, immigration and transmission. *Eur Respir J* 36(2):339–347.

92. Borgdorff MW, *et al*. (2011) The incubation period distribution of tuberculosis estimated with a molecular epidemiological approach. *Int J Epi* 40(4):964–970.

93. Cacho J, *et al.* (2007) Recurrent tuberculosis from 1992 to 2004 in a metropolitan area. *Eur Respir J* 30(2):333–337.

94. Crampin AC, *et al.* (2010) Recurrent TB: relapse or reinfection? The effect of HIV in a general population cohort in Malawi. *AIDS* 24(3):417–426.

95. Fleischmann RD, *et al*. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae*. *Science* 269(5223):496–512.

96. Metzker ML. (2009) Sequencing technologies — the next generation. *Nature Publishing Group* 11(1):31–46.

97. Loman NJ, *et al.* (2012) High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nat Rev Micro* 10(9):599–606.

98. Hasman H, *et al*. (2014) Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples. *J Clin Micro*

52(1):139–146.

99. Doughty EL, Sergeant MJ, Adetifa I, Antonio M, Pallen MJ. (2014) Culture-independent detection and characterisation of M*ycobacterium tuberculosis* and M. africanum in sputum samples using shotgun metagenomics on a benchtop sequencer. *PeerJ*. doi:10.7717/peerj.585/supp-3.

100. Comas I, *et al*. (2010) Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. *Nat Genet* 42(6):498–503.

101. Roetzer A, *et al*. (2013) Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiological study. *PLoS Med* 10(2):e1001387.

102. Felsenstein J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*.

103. Niemann S, *et al.* (2009) Genomic diversity among drug sensitive and multidrug resistant isolates of *Mycobacterium tuberculosis* with identical DNA fingerprints. *PLoS ONE* 4(10):e7407.

104. Schurch AC, *et al*. (2010) High-resolution typing by integration of genome sequencing data in a large tuberculosis cluster. *J Clin Micro* 48(9):3403–3406.

105. Walker TM, *et al*. (2013) Whole-genome sequencing to delineate M*ycobacterium tuberculosis* outbreaks: a retrospective observational study. *Lancet Infect Dis* 13(2):137–146.

106. Bryant JM, *et al*. (2013) Inferring patient to patient transmission of *Mycobacterium tuberculosis* from whole genome sequencing data. *BMC Infect Dis* 13:110–110.

107. Stucki D, *et al.* (2015) Tracking a tuberculosis outbreak over 21 years: strain-specific single-nucleotide polymorphism typing combined with targeted whole-genome sequencing. *J Infect Dis* 211(8):1306–1316.

108. Jamieson FB, *et al.* (2014) Whole-genome sequencing of the *Mycobacterium tuberculosis* Manila sublineage results in less clustering and better resolution than mycobacterial interspersed repetitive-unit-variable-number tandem-repeat (MIRU-VNTR) typing and spoligotyping. *J Clin Micro* 52(10):3795–3798.

109. Mehaffy C, *et al.* (2014) Marked microevolution of a unique *Mycobacterium tuberculosis* strain in 17 years of ongoing transmission in a high risk population. *PLoS ONE* 9(11):e112928.

110. Perez-Lago L, *et al.* (2015) Fast and low-cost decentralized surveillance of transmission of tuberculosis based on strain-specific PCRs tailored from whole genome sequencing data: a pilot study. *CMI* 21(3):249–249.

111. Comas I, *et al.* (2013) Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. *Nat Genet* 45(10):1176–1182.

112. Hirsh AE, Tsolaki AG, DeRiemer K, Feldman MW, Small PM. (2004) Stable association between strains of *Mycobacterium tuberculosis* and their human host populations. *Proc Natl Acad Sci USA* 101(14):4871–4876.

113. Phelan JE, *et al.* (2016) Recombination in pe/ppe genes contributes to genetic variation in *Mycobacterium tuberculosis* lineages. *BMC Genomics*:1–12.

114. Colangeli R, *et al*. (2014) Whole Genome Sequencing of *Mycobacterium tuberculosis* reveals slow growth and low mutation rates during latent infections in humans. *PLoS ONE* 9(3):e91024.

115. Kato-Maeda M, *et al*. (2013) Use of whole genome sequencing to determine the microevolution of *Mycobacterium tuberculosis* during an outbreak. *PLoS ONE* 8(3):e58235.

116. Pérez-Lago L, *et al.* (2014) Whole genome sequencing analysis of intrapatient microevolution in *Mycobacterium tuberculosis*: potential impact on the inference of tuberculosis transmission. *J Infect Dis* 209(1):98–108.

117. Bryant JM, *et al*. (2013) Whole-genome sequencing to establish relapse or re-infection with *Mycobacterium tuberculosis*: a retrospective observational study. *Lancet Respir Med* 1(10):786–792.

118. Glynn JR, Vynnycky E, Fine PE. (1999) Influence of sampling on estimates of clustering and recent transmission of *Mycobacterium tuberculosis* derived from DNA fingerprinting techniques. *Am J Epi* 149(4):366–371.

119. Borgdorff MW, van den Hof S, Kalisvaart N, Kremer K, van Soolingen D. (2011) Influence of sampling on clustering and associations with risk factors in the molecular epidemiology of tuberculosis. *Am J Epi* 174(2):243–251.

120. Murray M (2002) Sampling bias in the molecular epidemiology of tuberculosis. *Emerging Infect Dis* 8(4):363–369.

121. Braden CR, *et al.* (1997) Interpretation of restriction fragment length polymorphism analysis of *Mycobacterium tuberculosis* isolates from a state with a large rural population. *J Infect Dis* 175(6):1446–1452.

122. Nguyen D, *et al*. (2003) Widespread pyrazinamide-resistant *Mycobacterium tuberculosis* family in a low-incidence setting. *J Clin Micro* 41(7):2878–2883.

123. Guerra-Assuncao JA, et al. (2015) 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. *J Infect Dis* 211(7):1154–1163.

124. Eldholm V, *et al*. (2015) Four decades of transmission of a multidrug-resistant *Mycobacterium tuberculosis* outbreak strain. *Nature Commun* 6:1–9.

125. Cohen KA, *et al.* (2015) Evolution of extensively drug-resistant tuberculosis over four decades: whole genome sequencing and dating analysis of *Mycobacterium tuberculosis* isolates from KwaZulu-Natal. *PLoS Med* 12(9):e1001880.

126. Andersson DI, Hughes D. (2010) Antibiotic resistance and its cost: is it possible to reverse resistance? *Nat Rev Micro* 8(4):260–271.

127. Casali N, *et al*. (2014) Evolution and transmission of drug-resistant tuberculosis in a Russian population. *Nat Genet* 46(3):279–286.

128. Statistics Canada. *Nunavik, Inuit region, Quebec (Code 640002) Nunavik, Inuit region, Quebec (Code 640002).* Available at: http://www12.statcan.gc.ca/nhs-enm/2011/dp-pd/aprof/index.cfm?Lang=E.

129. Zar HJ, Hanslo D, Apolles P, Swingler G, Hussey G (2005) Induced sputum versus gastric lavage for microbiological confirmation of pulmonary tuberculosis in infants and young children: a prospective study. *Lancet* 365(9454):130–134.

130. Parrish NM, Carroll KC (2011) Role of the clinical mycobacteriology laboratory in diagnosis and management of tuberculosis in low-prevalence settings. *journal of clinical microbiology* 49(3):772–776.

131. van Soolingen D, Hermans PW, de Haas PE, Soll DR, van Embden JD (1991) 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. *J Clin Micro* 29(11):2578–2586.

132. Illumina (2012) TruSeq DNA Sample Preparation Guide. 1–148.

133. Illumina (2011) Paired-End Sample Preparation Guide. 1–40.

134. Illumina (2010) Illumina Sequencing Technology. 1–5.

135. Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* 36(16):e105–e105.

136. Pabinger S, *et al.* (2014) A survey of tools for variant analysis of next-generation genome sequencing data. *Brief Bioinform* 15(2):256–278.

137. Olson ND, *et al*. (2015) Best practices for evaluating single nucleotide variant calling methods for microbial genomics. *Front Genet* 6:235. doi:10.3389/fgene.2015.00235.

138. Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with

BWA-MEM. *arXiv*.

139. Thorvaldsdottir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 14(2):178–192.

140. Li H, et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078-2079.

141. McKenna A, *et al*. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20(9):1297–1303.

142. Pirooznia M, *et al*. (2014) Validation and assessment of variant calling pipelines for next-generation sequencing. *Hum Genomics* 8:14.

143. Liu X, Han S, Wang Z, Gelernter J, Yang B-Z (2013) Variant callers for next-generation sequencing data: a comparison study. *PLoS ONE* 8(9):e75619.

144. Eyre DW, *et al.* (2012) A pilot study of rapid benchtop sequencing of *Staphylococcus aureu*s and *Clostridium difficile f*or outbreak detection and surveillance. *BMJ Open* 2(3): e001124 doi:10.1136/bmjopen-2012-001124.

145. Harris KA, *et al.* (2015) Whole-genome sequencing and epidemiological analysis do not provide evidence for cross-transmission of *Mycobacterium abscessus* in a cohort of pediatric cystic fibrosis patients. *Clin Infect Dis* 60(7):1007–1016.

146. Eyre DW, *et al.* (2013) Detection of mixed infection from bacterial whole genome sequence data allows assessment of its role in *Clostridium difficile* transmission. *PLoS Comput Biol* 9(5):e1003059.

147. Danecek P, *et al.* (2011) The variant call format and VCFtools. *Bioinformatics* 27(15):2156–2158.

148. Cingolani P, *et al*. (2012) 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(2):80–92.

149. Meacham F, *et al.* (2011) Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinform* 12(1):451.

150. Sanger F, Nicklen S, Coulson AR. (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 74(12):5463–5467.

151. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990) Basic local alignment search tool. *J Mol Biol* 215(3):403–410.

152. Cohen T, *et al.* (2012) Mixed-strain *Mycobacterium tuberculosis* infections and the implications for tuberculosis treatment and control. *Clin Micro Rev* 25(4):708–719.

153. Felsenstein J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* 17(6):368–376.

154. Posada D, Crandall KA. (2001) Selecting models of nucleotide substitution: an application to human immunodeficiency virus 1 (HIV-1). *Mol Biol Evol* 18(6):897–906.

155. Huelsenbeck JP, Crandall KA. (1997) Phylogeny estimation and hypothesis testing using maximum likelihood. *Annu Rev Ecol Syst* 28:437–466.

156. Posada D, Crandall K. (2001) Simple (wrong) models for complex trees: a case from retroviridae. *Mol Biol Evol* 18(2):271–275.

157. Rzhetsky A, Sitnikova T. (1996) When is it safe to use an oversimplified substitution model in tree-making? *Mol Biol Evol* 13(9):1255–1265.

158. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. (2013) MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Mol Biol Evol* 30(12):2725–2729.

159. Waddell PJ, Steel MA. (1997) General time-reversible distances with unequal rates across sites: mixing gamma and inverse Gaussian distributions with invariant sites. *Mol Phylogenet Evol* 8(3):398–414.

160. Tamura K. (1992) Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Mol Biol Evol* 9(4):678–687.

161. Alfaro ME. (2003) Bayes or Bootstrap? A Simulation Study Comparing the Performance of Bayesian Markov Chain Monte Carlo Sampling and Bootstrapping in Assessing Phylogenetic Confidence. *Mol Biol Evol* 20(2):255–266.

162. Hillis DM, Bull JJ. (1993) An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology* 42(2): 182-192.

163. Reis dos M, Donoghue PCJ, Yang Z. (2015) Bayesian molecular clock dating of species divergences in the genomics era. *Nature Reviews Genetics* 17(2):71–80.

164. Drummond AJ, Bouckaert R. (2015) *Bayesian Evolutionary Analysis with BEAST* (Cambridge University Press).

165. Dunson DB. (2001) Commentary: practical advantages of Bayesian analysis of epidemiologic data. *Am J Epi* 153(12):1222–1226.

166. Drummond AJ, A SM, Xie D, Rambaut A. (2012) Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* 29(8):1969–1973.

167. Beiko R, Keith J, Harlow T, Ragan M. (2006) Searching for convergence in phylogenetic Markov Chain Monte Carlo. *Systematic Biology* 55(4):553–565.

168. Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W (2002) Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* 161(3):1307–1320.

169. Rambaut A, Suchard MA, Xie D, Drummond AJ Tracer. Available at: http://beast.bio.ed.ac.uk/Tracer.

170. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biol* 4(5):e88.

171. Baele G, *et al*. (2012) Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Mol Biol Evol* 29(9):2157–2167.

172. Drummond AJ. (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol* 22(5):1185–1192.

173. Rambaut A. (2014) Summarizing posterior trees. Available at: http://beast.bio.ed.ac.uk/summarizing-posterior-trees [Accessed November 25, 2014].

174. Drummond A, Rambaut A, Bouckaert R. (2013) Divergence dating tutorial with BEAST 2.0.

175. Bouckaert R, *et al*. (2014) BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput Biol* 10(4):e1003537.

176. Long SW, Beres SB, Olsen RJ, Musser JM. (2014) Absence of patient-to-patient intrahospital transmission of *Staphylococcus aureus* as determined by whole-genome sequencing. *mBio* 5(5):e01692–e01614.

177. Price JR, *et al*. (2014) Whole-genome sequencing shows that patient-to-patient transmission rarely accounts for acquisition of *Staphylococcus aureus* in an intensive care unit. *Clin Infect Dis* 58(5):609–618.

178. SenGupta DJ, *et al.* (2014) Whole-genome sequencing for high-resolution investigation of methicillin-resistant *Staphylococcus aureus* epidemiology and genome plasticity. *J Clin Micro* 52(8):2787–2796.

179. Snitkin ES, *et al.* (2012) Tracking a hospital outbreak of carbapenem-resistant *Klebsiella pneumoniae* with whole-genome sequencing. *Sci Transl Med* 4(148):148ra116–148ra116.

180. Quick J, *et al.* (2015) Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of *Salmonella*. *Genome Biol* 16(1):114–114.

181. Quick J, *et al.* (2014) Seeking the source of *Pseudomonas aeruginosa* infections in a recently opened hospital: an observational study using whole-genome sequencing.

BMJ Open (4):1–10.

182. Drobniewski F, *et al.* (2005) Drug-resistant tuberculosis, clinical virulence, and the dominance of the Beijing strain family in Russia. *JAMA* 293(22):2726–2731.

183. Dominguez J, *et al.* (2016) Clinical implications of molecular drug resistance testing for *Mycobacterium tuberculosis*: a TBNET/RESIST-TB consensus statement. *Int J Tuberc Lung Dis* 20(1):24–42.

184. Walker TM, *et al.* (2014) Assessment of *Mycobacterium tuberculosis* transmission in Oxfordshire, UK, 2007-12, with whole pathogen genome sequences: an observational study. *Lancet Respir Med* 2(4):285–292.

185. Guerra-Assuncao JA, Crampin AC, Houben R (2015) Large-scale whole genome sequencing of *M. tuberculosis* provides insights into transmission in a high prevalence area. *Elife*. doi:10.7554/eLife.05166.001.

186. Martin A, Herranz M, Serrano MJR, Bouza E, de Viedma DG. (2010) The clonal composition of Mycobacterium tuberculosis in clinical specimens could be modified by culture. *Tuberculosis* 90(3):201–207.

187. Brown AC, *et al.* (2015) Rapid Whole-Genome Sequencing of *Mycobacterium tuberculosis* isolates directly from clinical samples. *J Clin Micro* 53(7):2230–2237.

188. Black PA, *et al*. (2015) Whole genome sequencing reveals genomic heterogeneity and antibiotic purification in *Mycobacterium tuberculosis* isolates. *BMC Genomics*:1–14.

189. Ford C, *et al.* (2012) *Mycobacterium tuberculosis* – Heterogeneity revealed through whole genome sequencing. *Tuberculosis* 92(3):194–201.

190. Perez-Lago L, *et al.* (2015) Revealing hidden clonal complexity in *Mycobacterium tuberculosis* infection by qualitative and quantitative improvement of sampling. *CMI* 21(2):147.e1–147.e7.

191. Hatherell H-A, *et al.* (2016) Interpreting whole genome sequencing for investigating tuberculosis transmission: a systematic review. *BMC Med*:1–13.

192. Cohen T, *et al.* (2016) Within-host heterogeneity of *Mycobacterium tuberculosis* infection is associated with poor early treatment response: a prospective cohort Study. *J Infect Dis*: doi: 10.1093/infdis/jiw014.

APPENDIX 1

Glossary of terms

A1. Acronyms and glossary of terms

bp	Base-pairs (nucleotides)
BAM file	Binary Alignment Map file – this is a binary version of a Sequence Alignment Map (SAM) file, which stores alignments of nucleotide sequences in tab-delimited format (1).
Bioinformatics pipeline	Referring to the data analysis steps for WGS; specifically, the steps to convert raw sequence reads to a dataset of single nucleotide polymorphisms. The precise steps and software used for each vary between WGS studies. Each unique combination of steps and software represents a unique pipeline.
CLSC	Centre Locale de Services Communautaires – Provides outpatient and emergency clinical services (in Nunavik, these are also called 'nursing stations').
Depth of coverage	Refers to the number of times a particular locus in the genome has been sequenced.
dN/dS	The ratio of nonsynonymous to synonymous single nucleotide polymorphisms, used to evaluate evolutionary pressure and selection on proteins across a population (2).
DST	Drug susceptibility testing.
ESS	Effective sampling size – Represents the number of independent samples that can be drawn from the posterior distribution. This is less than the total number of samples drawn, as adjacent samples are highly correlated.
Genome coverage	The percentage of the reference with at least X reads aligned to it (most studies report this with X=1).
HPD interval	Highest posterior density interval – represents the narrowest credible interval containing X% of the posterior probability (3)
Indels	Insertions and deletions.
Isolates	The bacteria obtained from individuals with micro-biologically confirmed active tuberculosis.
LTBI	Latent tuberculosis infection - When exposed to an active TB case, some individuals become infected. Individuals who do not progress to active TB within the first 2 years are considered to have LTBI.
MCMC	Markov chain Monte Carlo method.
MDR-TB	Multi-drug resistant TB - Resistance to at least isoniazid and rifampin, two of the front-line anti-tuberculosis drugs.
Micro-evolution	As *M. tuberculosis* replicates within the host, some bacteria acquire mutations. This can lead to a diversity of strains within the same patient, some, which have a mutation in a particular locus, and others without, all arising from the same initial strain.
MRCA	Most recent common ancestor – this represents an inferred ancestor of 2 isolates. The sequence of this ancestor can be reconstructed. 'tMRCA' represents the time at which the isolates diverged from this ancestor.

MIRU	Mycobacterial interspersed repetitive units - Evaluates the number of repetitive units of a particular locus in the MTB genome. Current recommendation is 24 loci, with a minimum of 15.
MUHC	McGill University Health Centre.
New positive TST	When an individual has a positive tuberculin skin test, with no previous test documented. Therefore, infection could have occurred at any time in the individual's life thus far.
NGS	Next-generation sequencing – High-throughput DNA sequencing, where thousands of sequences are produced simultaneously.
Non-synonymous SNP	This is a change in a single base that results in a change at the amino acid, and consequently, protein level. These are often deleterious.
NRBHSS	Nunavik Regional Board of Health and Social Services.
PCR	Polymerase chain reaction.
PE_PGRS and PPE genes	Proline-glutamate and proline-proline-glutamate gene families – these represent 10% of the coding genes in *M. tuberculosis* (4) and are highly repetitive, therefore the current recommendation is to exclude SNPs in these regions from analysis when short-read data is used (5, 6), due to probable mapping errors.
Phred score	$-10*\log P_{error}$; reflects the probability of error in a given base call, or when applied to mapping, the error in alignment.
PPV	Positive predictive value – P(D+|T+).
Reads	These are the sequences of each DNA fragment, and are used to re-construct the genome of the isolate under investigation.
Reference-based assembly	Alignment ('mapping') of reads to the correct locus according to a reference genome.
RFLP	Restriction fragment length polymorphism - Examines the location of certain insertion sequences in the MTB genome.
Sanger Sequencing	This method was invented in 1977 (7) and was previously the gold standard for DNA sequencing. DNA polymerase is used to copy denatured DNA by incorporating nucleotides from 5' to 3'. Specifically, the single-stranded DNA is mixed with a radioactive primer, each nucleotide (A, T, C, G), and a small percentage of a DNA analogue of one of these nucleotides. Whenever this analogue (which lacks a 3' hydroxyl group) is incorporated instead of the corresponding nucleotide, sequencing of that particular strand of DNA stops. Once the reaction is complete, all DNA sequences are separated by size using gel electrophoresis and the position of the analogue can be inferred. This is repeated using an analogue of each nucleotide, to identify the overall sequence.
Sequencing by synthesis	In brief, DNA polymerase is used to add synthesize DNA, copying the complementary strand by adding nucleotides from 5' to 3'. Unlike Sanger sequencing, as nucleotides are identified as they are incorporated, i.e. synthesis of the DNA and sequence identification are performed simultaneously.

Short-reads	Reads of up to 300 base-pairs in length (by Illumina MiSeq). Optimal for reference-based assembly, i.e. alignment using a reference genome.
Smear-positive	Smear positive for acid-fast bacilli on microscopy. Grades are assigned based on the concentration of such bacilli per unit area (1+, 2+, 3+, 4+), with higher numbers reflecting higher bacterial load. (8)
SNP	Single nucleotide polymorphism - Change in a single base of the genome, compared to a reference
SNP calling	The process of detecting whether or not each base in the genome is a single nucleotide polymorphism (SNP), i.e. whether the base is different from that found at the same position in the reference genome. Different algorithms can be used to identify ('call') SNPs.
Synonymous SNP	A change in a single base that does not result in a change in the amino acid. Through regulatory effects, however, these may influence expression of protein levels.
TST	Tuberculin skin test.
TST conversion	When an individual has a positive tuberculin skin test, following a documented negative test. This suggests infection occurred at some point in time between the two tests.
WGS	Whole genome sequencing - Sequencing the entire genome of an organism.
XDR-TB	Extremely drug resistant tuberculosis - Resistance to at least isoniazid and rifampin, two of the front-line anti-tuberculosis drugs, fluoroquinolones and at least one an injectable.

References

1. Li H, *et al*. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078–2079.

2. Kryazhimskiy S, Plotkin JB (2008) The population genetics of dN/dS. *PLoS Genet* 4(12):e1000304.

3. Drummond AJ, Bouckaert R (2015) *Bayesian Evolutionary Analysis with BEAST* (Cambridge University Press).

4. Cole ST, *et al*. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393(6685):537–544.

5. Roetzer A, *et al*. (2013) Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiological study. *PLoS Med* 10(2):e1001387.

6. Comas I, *et al.* (2010) Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. *Nat Genet* 42(6):498–503.

7. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 74(12):5463–5467.

8. Public Health Agency of Canada, Canadian Lung Association, Canadian Thoracic Society (2014) *Canadian Tuberculosis Standards 7th Edition*.

APPENDIX 2-1

Reprint of:

Lee RS, Radomski N, Proulx J-F, Manry J, McIntosh F, Desjardins F, Soualhine H, Domenech P, Reed MB, Menzies D, Behr MA. Reemergence and re-amplification of tuberculosis in the Canadian Arctic. *J Infect Dis* 2015;211(12):1905-1914

Reemergence and Amplification of Tuberculosis in the Canadian Arctic

Robyn S. Lee,[1,4,5,a] Nicolas Radomski,[5,a] Jean-Francois Proulx,[8] Jeremy Manry,[2,3,4,5] Fiona McIntosh,[5] Francine Desjardins,[6] Hafid Soualhine,[9] Pilar Domenech,[5] Michael B. Reed,[4,5] Dick Menzies,[5,7] and Marcel A. Behr[4,5]

Departments of [1]Epidemiology, Biostatistics and Occupational Health, [2]Medicine, and [3]Human Genetics, McGill University, [4]McGill International TB Centre, [5]The Research Institute of the McGill University Health Centre, [6]Mycobacteriology Laboratory, Royal Victoria Hospital, and [7]Respiratory Epidemiology and Clinical Research Unit, Montreal Chest Institute, McGill University Health Centre, [8]Nunavik Regional Board of Health and Social Services, Kuujjuaq, and [9]Laboratoire de Santé Publique du Québec, Sainte-Anne-de-Bellevue, Québec, Canada

Background. Between November 2011 and November 2012, a Canadian village of 933 persons had 50 culture-positive cases of tuberculosis, with 49 sharing the same genotype.

Methods. We performed Illumina-based whole-genome sequencing on *Mycobacterium tuberculosis* isolates from this village, during and before the outbreak. Phylogenetic trees were generated using the maximum likelihood method.

Results. Three distinct genotypes were identified. Strain I (n = 7) was isolated in 1991–1996. Strain II (n = 8) was isolated in 1996–2004. Strain III (n = 62) first appeared in 2007 and did not arise from strain I or II. Within strain III, there were 3 related but distinct clusters: IIIA, IIIB, and IIIC. Between 2007 and 2010, cluster IIIA predominated (11 of 22 vs 2 of 40; $P < .001$), whereas in 2011–2012 clusters IIIB (n = 18) and IIIC (n = 20) predominated over cluster IIIA (n = 11). Combined evolutionary and epidemiologic analysis of strain III cases revealed that the outbreak in 2011–2012 was the result of ≥6 temporally staggered events, spanning from 1 reactivation case to a point-source outbreak of 20 cases.

Conclusions. After the disappearance of 2 strains of *M. tuberculosis* in this village, its reemergence in 2007 was followed by an epidemiologic amplification, affecting >5% of the population.

Keywords. infectious disease outbreaks; *Mycobacterium tuberculosis*; molecular epidemiology; whole genome sequencing; transmission.

Between November 2011 and November 2012, there were 50 cases of microbiologically proven active tuberculosis in an Arctic village in Nunavik, Québec. With a population of only 933, the incidence of culture-confirmed tuberculosis was >5% of the community for that year—1000 times the overall Canadian incidence. This outbreak occurred in a setting with a very low prevalence of human immunodeficiency virus infection and no previous resistance to antituberculosis drugs, leading to concern in the populace of a newly emerged hypervirulent strain of *Mycobacterium tuberculosis*.

As part of the response to the outbreak, the Nunavik Regional Board of Health and Social Services (NRBHSS) conducted extensive contact investigations of all newly diagnosed active tuberculosis cases, including household and social contacts. During this response, it was observed that many persons had contacts with multiple tuberculosis cases, indicating that it would be extremely difficult to identify transmission links using standard epidemiologic methods. An alternative approach would involve molecular typing of patient isolates.

In work published elsewhere, a combination of classic molecular epidemiology tools (restriction fragment length polymorphism [RFLP] and mycobacterial interspersed repetitive units [MIRUs]) revealed extremely limited bacterial diversity in this region, both within and across villages [1]. One potential interpretation of these findings is that this represents ongoing transmission. However, an alternative hypothesis is that patients share similar bacterial genotypes due to ancestry. With the advent of whole-genome sequencing (WGS), a higher-resolution molecular epidemiologic

Received 11 October 2014; accepted 19 December 2014; electronically published 9 January 2015.

[a]R. S. L. and N. R. contributed equally to this work.

Correspondence: Marcel A. Behr, MD, MSc, McGill University Health Centre, Montreal General Hospital 1650 Cedar Ave, Room A5.156 Montreal, Québec, Canada H3G 1A4 (marcel.behr@mcgill.ca).

The Journal of Infectious Diseases® 2015;211:1905–14

DOI: 10.1093/infdis/jiv011

tool [2–6], it is now possible to test whether bacteria that are otherwise indistinguishable indicate recent transmission of *M. tuberculosis*. Furthermore, because WGS provides information on lineage-specific polymorphisms, this genotyping method can also determine whether a new, potentially more virulent *M. tuberculosis* strain had been introduced into this community.

To address these 2 questions, we conducted WGS on *M. tuberculosis* isolates from this village. To validate WGS data in this setting, we tested epidemiologically unrelated isolates from other villages of the same region, over 6 years. Then, to situate WGS data from 2011 to 2012 in the context of a village with high rates of tuberculosis over many years, we extended our analysis to the 2 decades before the outbreak. In this setting with limited variability by conventional genotyping modalities, WGS provided improved analytic resolution, revealing the disappearance, reemergence, and amplification of *M. tuberculosis* over time.

METHODS

Study Population

Nunavik, the arctic region of Québec, spans 443 685 km^2 and comprises 14 Inuit communities. The outbreak village, henceforth denoted village K, is >150 km from the nearest village, with no road connecting the communities.

Bacteria

Specimens from tuberculosis suspects in Nunavik are processed at the mycobacteriology laboratory of the McGill University Health Centre (MUHC). Culture-positive isolates are forwarded to the reference laboratory, Laboratoire de Santé Publique du Québec, for drug susceptibility testing. These laboratories provided isolates for the years 1991–2012.

Genomics

DNA extraction [7] and WGS have been described elsewhere [8], with details in the Supplementary Data. In brief, *M. tuberculosis* isolates were sequenced using the MiSeq 250 System (Illumina). Readings with a minimum length of 50 base pairs (bp) were retained and deposited in the National Center for Biotechnology Information's Sequence Archive (accession No. SRP039605, i.e. BioProject PRJNA240330). After alignment to the H36Rv reference genome (accession No. NC_000962.3), single-nucleotide polymorphisms (SNPs) were identified using a Bayesian likelihood model (Unified Genotyper; Genome Analysis Toolkit, version 2.7.4); SNPs with a minimum Phred score >50 were retained (where Phred is $-10 \cdot \log_{10} P_{error}$). Phylogenetic analysis was done using Molecular Evolutionary Genetics Analysis (MEGA, version 5, [9]), with the number of differences method used to compute evolutionary distance [10]. Maximum likelihood trees were generated using the model of nucleotide substitution that yielded the lowest Bayesian information criterion (Tamura 3-parameter model, [11]). As a sensitivity analysis, we also generated maximum likelihood trees using the Jukes–Cantor model [12].

Validation of SNP Threshold for Recent Transmission

Given the limited genetic diversity in Nunavik [1], we evaluated the lowest SNP threshold that could occur in the absence of transmission. To do so, we sequenced *M. tuberculosis* isolates from cases residing in other villages of Nunavik (2006–2012). Contact investigation data were obtained from the NRBHSS. Case pairs without epidemiologic links who resided in *different* villages were designated as *improbable transmission,* and the SNPs between these case pairs were compared.

Application of WGS to Village K

The SNPs between village K isolates were identified, including those from cases diagnosed during the 20 years before the outbreak. Phylogenetic trees were generated while blinded to epidemiologic data.

Clinical Epidemiologic Analysis Combined With WGS

For the outbreak, clinical epidemiologic data were collected by clinical staff in village K. Links between cases were identified using a database of all household and named contacts. Using date of diagnosis/treatment initiation, symptoms, sputum smear status, and cavity on chest radiograph as indicators of contagion [13], we looked for potential index cases in each cluster. For the years preceding the outbreak, epidemiologic data for cases from 2007 to 2010 were provided by the NRBHSS. Smear microscopic results were obtained from the MUHC laboratory.

Statistical Analysis

A 2-sample z test and the exact binomial test were used to compare proportions. The F* test for samples with unequal variance was used to compare the number of pairwise SNPs within clusters. Analyses were conducted using Stata software (version 11, StataCorp 2009).

Ethical Approval

Ethical approval was obtained from the McGill University Faculty of Medicine's institutional review board and the NRBHSS. Individual patient consent was not required, but the study was done in collaboration with the village K council.

RESULTS

The Outbreak

Between November 2011 and November 2012, there were 50 microbiologically confirmed cases of tuberculosis in village K. There were no cases between January and October of 2011. All cases were pulmonary, with no instances of tuberculosis meningitis or disseminated disease. Seven of the 50 cases were

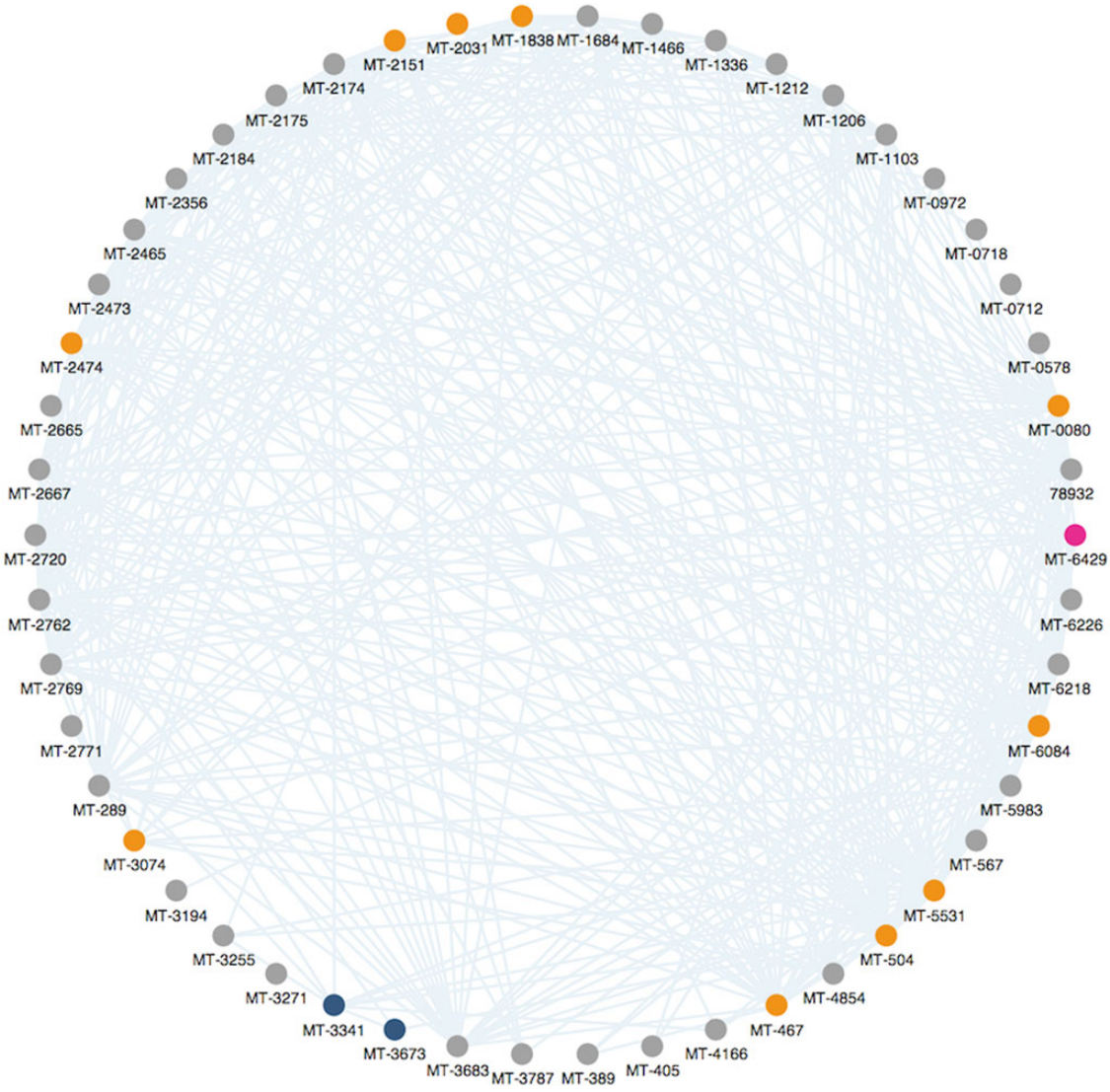


Figure 1. Epidemiologic links between outbreak cases. Links between household/named contacts, as well as shared attendance (or residence) of community "gathering houses" identified by contact investigation are indicated. Orange circles represent sputum smear-positive cavitary cases; navy circles, sputum smear-positive noncavitary cases; pink circles, sputum smear-negative cavitary cases; gray circles, sputum smear-negative noncavitary cases.

diagnosed based on symptoms. Of the remaining 43 cases, 40 were found to have active disease during contact investigation, and 3 developed tuberculosis after a documented positive tuberculin skin test conversion; 1 had refused isoniazid and the other 2 demonstrated low adherence. The epidemiologic links between cases were highly complex (Figure 1). All cases except one shared the same MIRU pattern; RFLP provided similar resolution (Supplementary Figure 1).

Tuberculosis in Village K Over 22 Years

Between 1991 and 2012 (ie, including the outbreak year), 82 cases of culture-positive tuberculosis were diagnosed in village K (Figure 2), yielding an average annual incidence of >450 per 100 000 (population denominators from Statistics Canada). The majority of cases were male (47 of 82), with a median age of 22 years (interquartile range, 16–35 years), consistent with the age distribution of this population [14].

Of the 82 confirmed cases in village K, 80 (97.6%) had isolates available for genotyping, 78 of which provided high-quality WGS data: 49 of 50 outbreak isolates, 14 of 15 isolates from 2007 to 2010, and all 15 isolates from 1991 to 2004 (there were no cases in 2005–2006). Average genome coverage among the 78 isolates was 99.7% (standard deviation [SD], 0.11%), with an average depth of coverage of 42× (SD, 13). The majority of Phred scores were between 500 and 1000 for SNPs, indicating minimal ascertainment bias, and there was no evidence

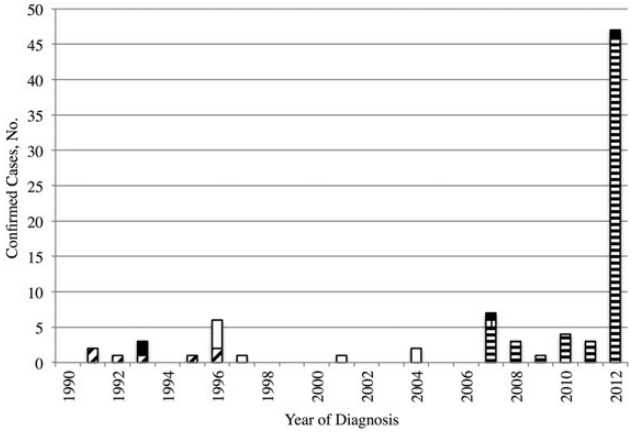


Figure 2. Microbiologically confirmed tuberculosis in village K (1990–2012). The numbers of confirmed tuberculosis cases reported in village K from 1990 to 2012 are shown by year of diagnosis. Strains of isolates are indicated, as identified by whole-genome sequencing (WGS): diagonal stripes indicate strain I; solid white, strain II; horizontal stripes, strain III; and vertical stripes, not clustered; solid black represent isolates for which WGS was not available.

supporting infection with multiple *M. tuberculosis* strains (Supplementary Figure 2). A review of isolates without SNPs between them revealed that the specimens were processed in separate batches, arguing against laboratory cross-contamination.

Validation of SNP Threshold for Recent Transmission

WGS was successful for 42 of 45 cases in other villages of Nunavik (2006–2012). Consistent with our observation of limited genetic diversity in this region, the 631 "improbable transmission" case pairs from other villages of Nunavik were separated by as few as 2 SNPs, but none were separated by 0 or 1 SNP (Supplementary Figure 3). From this finding, supported by studies published elsewhere, we defined a new cluster when a group of isolates shared ≥2 of the same SNPs compared with the reference group.

Application of WGS to Village K

The SNPs from all isolates of village K were used to infer maximum likelihood trees, with the bootstrap consensus tree from 1000 replicates shown in Figure 3 [11, 15]. Results were robust to use of an alternate model of nucleotide substitution (unpublished data). All isolates were lineage 4 (Euro-American, with the reported 7-bp deletion in the *pks15/1* gene) [16], and 3 distinct strains were evident, designated strains I, II, and III (Figure 3). Neither strain I nor strain II gave rise to strain III; strain I has 16 unique SNPs not seen in strain III, whereas strain II has 18 unique SNPs plus a 1102-bp deletion (2 963 340–2 964 352) that is intact in strain III isolates.

Strain I predominated for 6 years (n = 7; 1991–1996), then disappeared. Strain II predominated for 9 years (n = 8; 1996–2004), then disappeared (Figure 2). Strains I and II were unique to village K. Strain III was first detected in village K in 2007, though it was subsequently found in 2 cases diagnosed in other villages. One of these cases was a child adopted from village K to another community, and the other was an adult who had been a close family contact of a smear-positive case in village K before developing active tuberculosis the following year.

Within strain III, 3 clusters were observed, designated IIIA, IIIB, and IIIC (Figure 4). Cluster IIIA isolates (n = 22) had the reference alleles for the genes *carB, Rv3263*, *Rv0828c,* and *Rv1835c*. Cluster IIIB isolates (n = 20) had cluster-defining SNPs in *carB* and *Rv3263* but were wild type for *Rv0828c* and *Rv1835c;* cluster IIIC isolates (n = 20) had cluster-defining SNPs in *Rv0828c* and *Rv1835c* but were wild-type for *carB* and *Rv3263*. Of the 3 clusters, IIIC had the least bacterial diversity (mean pairwise SNP difference between isolates, 1.7 [95% confidence interval, 1.5–1.8] within IIIA, 1.6 (1.4–1.8) within IIIB, and 0.4 (0.3–0.5) within IIIC; $P < .001$).

Clinical Epidemiologic Analysis Combined With WGS

Whereas WGS alone revealed 3 different clusters (IIIA, B, C), further analysis in conjunction with epidemiologic data identified more complex transmission networks over time, with ≥6 distinct subgroups from 2011 to 2012 (Figure 5, across the bottom). Cluster IIIA was first seen in 2007–2008 and was initially divided into 2 groups—those with the C allele in *mce1B* (n = 4) and those with an alternative T allele in this gene (n = 18).

Between 2011 and 2012, there were 11 cluster IIIA isolates. One had the C allele in *mce1B* and was from a familial contact of previous cases whose isolates had the same genotype in 2008, suggestive of an isolated reactivation event. The 10 remaining isolates had the T allele in *mce1B.* Two of these isolates also had an alternative C allele in *Rv0331*. In this latter subgroup, 1 case was diagnosed in November 2011 and had smear-positive (3+) cavitary disease (MT-5531), while the other was a household contact. The remaining 8 IIIA isolates were first observed in May 2012. Within this subgroup, there were 3 smear-positive cases (4+ for MT-3074, 3+ for MT-3341, and 2+ for MT-3673) diagnosed in June 2012 plus 5 more cases diagnosed at about the same time or soon afterward. Nearly all secondary cases were friends or family, with no obvious trend in locations of contact. Thus, the 11 cluster IIIA isolates from 2011 to 2012 are unlikely to represent a single transmission event, because ≥2 discrete transmission chains plus 1 isolated reactivation event are better supported by the combined genetic and epidemiologic data.

Cluster IIIB was first seen in 2009 and had the reference *mce1B* C allele, plus cluster-defining SNPs in *carB* and *Rv3263*. In 2011–2012, there were 18 cluster IIIB isolates. Five of these had an alternative C allele in *fadE4,* and the other 13 had the reference A allele at this position. The former subgroup was

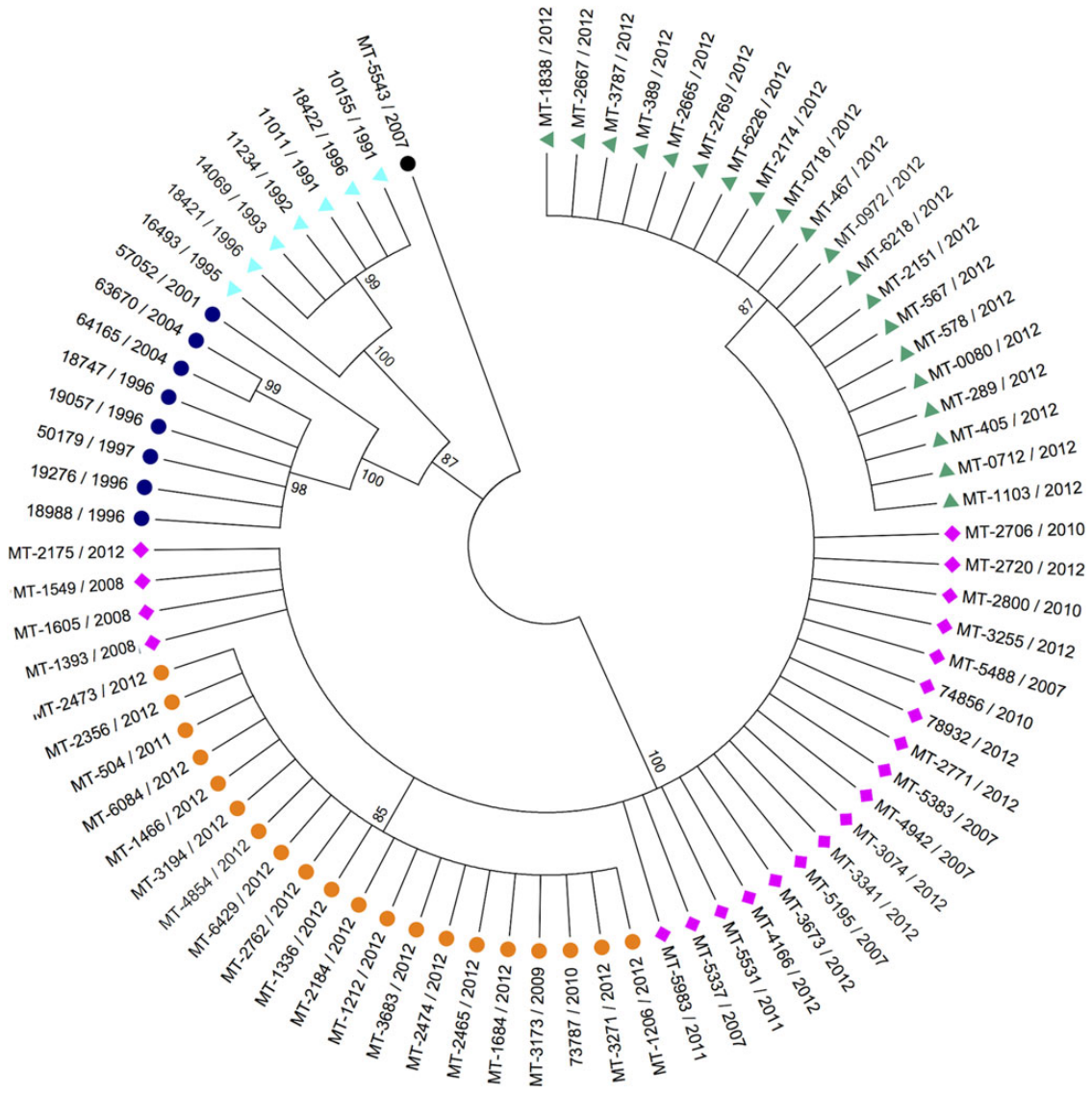


Figure 3. Bootstrap consensus tree of *Mycobacterium tuberculosis* isolates from village K. The evolutionary history was inferred by using the maximum likelihood method based on the Tamura 3-parameter model [11] and a bootstrap consensus tree was generated with 1000 replicates [15]. The percentage of trees in which the associated genome clustered together is shown next to the branches. Branches reproduced in <80% of bootstrap replicates are collapsed. Initial trees for the heuristic search were obtained by applying the neighbor-joining method to a matrix of pairwise distances estimated using the maximum composite likelihood approach. The analysis involved 78 genomes compared with the H37Rv reference genome. Light blue triangles represent strain I isolates; dark blue circles, strain II; pink diamonds, strain III, cluster A; orange circles, strain III, cluster B; green triangles, strain III, cluster C; black circle, not clustered.

first seen in December 2011, when a single case was diagnosed with smear-positive (4+) cavitary disease (MT-504). The remaining 4 cases with this genotype were teenagers with shared attendance at the same "gathering house," a venue of socialization identified by public health during the outbreak. The latter subgroup (with the reference A allele in *fadE4*) was first detected 3 months later, in March 2012. Although it is possible that MT-504 had a mixed infection and contributed to both subgroups, we also note that cases with the alternative C allele were diagnosed months before those with the reference A allele. Moreover, the group of 13 cases with the reference A allele included a patient with smear-positive (3+) cavitary disease diagnosed in May 2012 (MT-2474) who had definitive epidemiologic links to 4 of the remaining 12 cases. The combination of WGS and epidemiology together suggest that the 18 cluster IIIB isolates from 2011 to 2012 represent ≥2 transmission chains.

Cluster IIIC was not seen in the community before 2012. The first case was diagnosed in January 2012 with sputum smear–positive (3+) cavitary disease (MT-0080). Fifteen of the remaining 19 cases were epidemiologically linked to this case (4 household contacts, 3 friends, and 8 contacts at gathering houses). This putative source reported symptoms for 4 months before diagnosis,

						Strain				
						I	II	III		
								Cluster		
Position in H37Rv	Gene code	Gene name	Reference allele	Alternative allele	Mutation			IIIA	IIIB	IIIC
1331500	Rv1188	*Rv1188*	A	G	Nonsynonymous	G	G	A	A	A
1567583	Rv1392	*metK*	C	T	Synonymous	T	T	C	C	C
2022484	Rv1784	*Rv1784*	C	T	Synonymous	T	T	C	C	C
22398	Rv0018c	*ppp*	T	G	Nonsynonymous	G	T	T	T	T
42401	N/A	N/A	G	C	Intergenic	C	G	G	G	G
383629	Rv0315	*Rv0315*	C	T	Nonsynonymous	T	C	C	C	C
428744	Rv0355c	*PPE8*	A	G	Nonsynonymous	G	A	A	A	A
558829	N/A	N/A	C	G	Intergenic	G	C	C	C	C
1011972	Rv0908	*ctpE*	G	A	Nonsynonymous	A	G	G	G	G
2115139	Rv1866	*Rv1866*	G	A	Nonsynonymous	A	G	G	G	G
3251386	Rv2932	*ppsB*	C	T	Synonymous	T	C	C	C	C
3280352	Rv2940c	*mas*	C	A	Nonsynonymous	A	C	C	C	C
3367489	Rv3009c	*gatB*	T	C	Synonymous	C	T	T	T	T
3368965	Rv3010c	*pfkA*	T	G	Nonsynonymous	G	T	T	T	T
3910540	Rv3492c	*Rv3492c*	G	A	Synonymous	A	G	G	G	G
4093885	Rv3652	*PE_PGRS60*	C	T	Nonsynonymous	T	C	C	C	C
30642	Rv0026	*Rv0026*	C	T	Synonymous	C	T	C	C	C
403587	Rv0338c	*Rv0338c*	G	A	Nonsynonymous	G	A	G	G	G
677113	N/A	N/A	C	T	Intergenic	C	T	C	C	C
775381	Rv0675	*echA5*	A	C	Nonsynonymous	A	C	A	A	A
852682	Rv0758	*phoR*	T	G	Nonsynonymous	T	G	T	T	T
1310662	Rv1179c	*Rv1179c*	C	T	Nonsynonymous	C	T	C	C	C
2472939	Rv2208	*cobS*	A	C	Synonymous	A	C	A	A	A
2706379	Rv2408	*PE24*	C	T	Synonymous	C	T	C	C	C
3137638	Rv2831	*echA16*	C	G	Nonsynonymous	C	G	C	C	C
3206398	Rv2896c	*Rv2896c*	A	C	Nonsynonymous	A	C	A	A	A
3503781	Rv3137	*Rv3137*	G	A	Nonsynonymous	G	A	G	G	G
3973994	Rv3535c	*Rv3535c*	C	T	Synonymous	C	T	C	C	C
3975488	Rv3537	*Rv3537*	C	T	Synonymous	C	T	C	C	C
3997479	Rv3557c	*Rv3557c*	G	A	Synonymous	G	A	G	G	G
4218303	Rv3773c	*Rv3773c*	G	A	Nonsynonymous	G	A	G	G	G
567913	N/A	N/A	T	C	Intergenic	T	T	C	C	C
600950	Rv0509	*hemA*	G	A	Synonymous	G	G	A	A	A
809416	Rv0712	*Rv0712*	A	G	Nonsynonymous	A	A	G	G	G
1360496	Rv1217c	*Rv1217c*	C	A	Nonsynonymous	C	C	A	A	A
1837767	Rv1633	*uvrB*	G	A	Synonymous	G	G	A	A	A
1917190	Rv1692	*Rv1692*	G	A	Nonsynonymous	G	G	A	A	A
1973649	Rv1747	*Rv1747*	C	T	Nonsynonymous	C	C	T	T	T
2228900	N/A	N/A	T	G	Intergenic	T	T	G	G	G
2501354	Rv2227	*Rv2227*	G	A	Nonsynonymous	G	G	A	A	A
2623818	N/A	N/A	G	A	Intergenic	G	G	A	A	A
2760387	Rv2458	*mmuM*	A	C	Synonymous	A	A	C	C	C
2783792	Rv2477c	*Rv2477c*	G	T	Nonsynonymous	G	G	T	T	T
2976814	Rv2653c	*Rv2653c*	G	A	Synonymous	G	G	A	A	A
3108453	Rv2800	*Rv2800*	T	C	Nonsynonymous	T	T	C	C	C
3374067	Rv3014c	*ligA*	A	C	Nonsynonymous	A	A	C	C	C
3430937	Rv3066	*Rv3066*	G	A	Synonymous	G	G	A	A	A
3467000	Rv3097c	*lipY*	T	A	Nonsynonymous	T	T	A	A	A
3599965	Rv3224	*Rv3224*	G	A	Nonsynonymous	G	G	A	A	A
3772940	Rv3360	*Rv3360*	C	T	Nonsynonymous	C	C	T	T	T
4232293	Rv3785	*Rv3785*	C	G	Nonsynonymous	C	C	G	G	G
4287890	Rv3822	*Rv3822*	G	A	Synonymous	G	G	A	A	A
4409769	Rv3921c	*Rv3921c*	G	C	Nonsynonymous	G	G	C	C	C
1558108	Rv1384	*carB*	C	T	Synonymous	C	C	C	T	C
3644579	Rv3263	*Rv3263*	G	A	Nonsynonymous	G	G	G	A	G
921390	Rv0828c	*Rv0828c*	T	C	Nonsynonymous	T	T	T	T	C
2082436	Rv1835c	*Rv1835c*	C	T	Nonsynonymous	C	C	C	C	T

Figure 4. Strain and cluster-defining single-nucleotide polymorphisms (SNPs) for strains I, II, and III. Strain and cluster-defining SNPs shown. Reference and alternative alleles are highlighted in white and gray, respectively. From a progenitor strain, strains I and II have evolved distinctly from strain III, itself further subdivided into clusters IIIA, IIIB, and IIIC. Alleles in the genes *Rv0828c*, *carB*, *Rv1835c*, and *Rv3263* (H37Rv loci 1 558 108, 3 644 579, 921 390 and 2 082 436, respectively) were confirmed by Sanger sequencing for 6 isolates from each of clusters IIIA, IIIB, and IIIC.

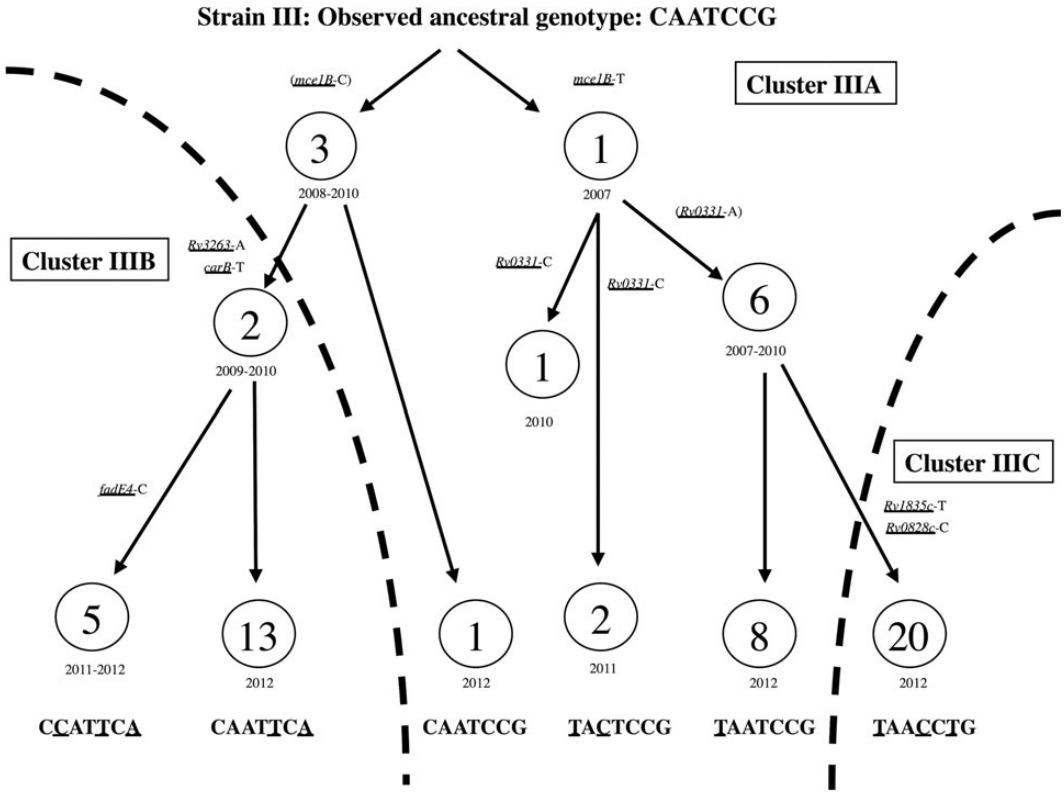


Figure 5. The microevolution of strain III in village K over time, involving a total of 7 single-nucleotide polymorphisms (SNPs). Numbers in circles indicate numbers of cases at each stage of evolution. The years of all isolates in each group are indicated below the circles, with time scaled from the top (2007) to the bottom (2012). Arrows indicate bacterial microevolution, and SNPs are identified by the gene name and the corresponding allele; to highlight certain lineages with the reference allele, the gene name and allele are in parentheses. Starting with the ancestral genome (*top*), cluster IIIA had 2 initial subgroups, one with the reference allele C at *mce1B* (4 cases; *middle panel, left*) and the other with alternative allele T at *mce1B* (18 cases; *middle panel, right*). Within the latter 18 cases, there were 2 subgroups: 3 with an additional variant at *Rv0331* and 15 that retained the reference allele, A. Cluster IIIB (*bottom left*) was derived from strain IIIA with the *mce1B* C reference allele, with 2 additional mutations (in *Rv3263*, *carB*). A subgroup of 5 had an additional variant in *fadE4*. Cluster IIIC (*bottom right*) was derived from cluster IIIA, with the alternative T allele at *mce1B* and 2 additional mutations (in *Rv1835c*, *Rv0828c*). At the bottom, the concatenated genotype for the 7 SNPs is presented for each of the 6 subgroups identified during the outbreak year.

possibly explaining the large number of IIIC cases observed early in 2012 (8 additional cases in January–February 2012 and 3 in March–April). Of these cases, 2 were smear positive (2+ for MT-1838 and 2+ for MT-2151). Hence, some of the remaining cases with diagnoses between May and November 2012 may have been infected by these secondary cases. These data suggest that cluster IIIC represents, at a minimum, 1 discrete transmission chain.

The epidemiologic curve of the outbreak shows, at the village level, a bimodal distribution of cases diagnosed over time (Figure 6*A*). When outbreak cases were stratified by the aforementioned subgroups, the bimodal distribution was largely attributable to differences in the temporal presentation of the different clusters and their subgroups (Figure 6*B*). When examining the contact data on the most transmissible cases in each of the subgroups, we can tabulate the number of household and nonhousehold contacts who developed active tuberculosis with the same genotype. As seen in Table 1, of named household contacts who developed tuberculosis, 56% shared the same genotype as the epidemiologically identified source. In contrast, among nonhousehold contacts who developed tuberculosis, only 19% shared the same genotype as their putative source, which was no better than chance alone (exact binomial for comparison to 1/6, given 6 subgroups; $P = .32$).

DISCUSSION

Using WGS, we have been able to reveal the complexity of tuberculosis control in a unique environment, where there is virtually no loss to follow-up and little to no in- or out-migration. On the scale of decades, 2 dominant strains have disappeared, not to be seen again after 1996 and 2004. Unfortunately, the reemergence of tuberculosis in or around 2007 was followed by a series of secondary and tertiary cases, culminating in an

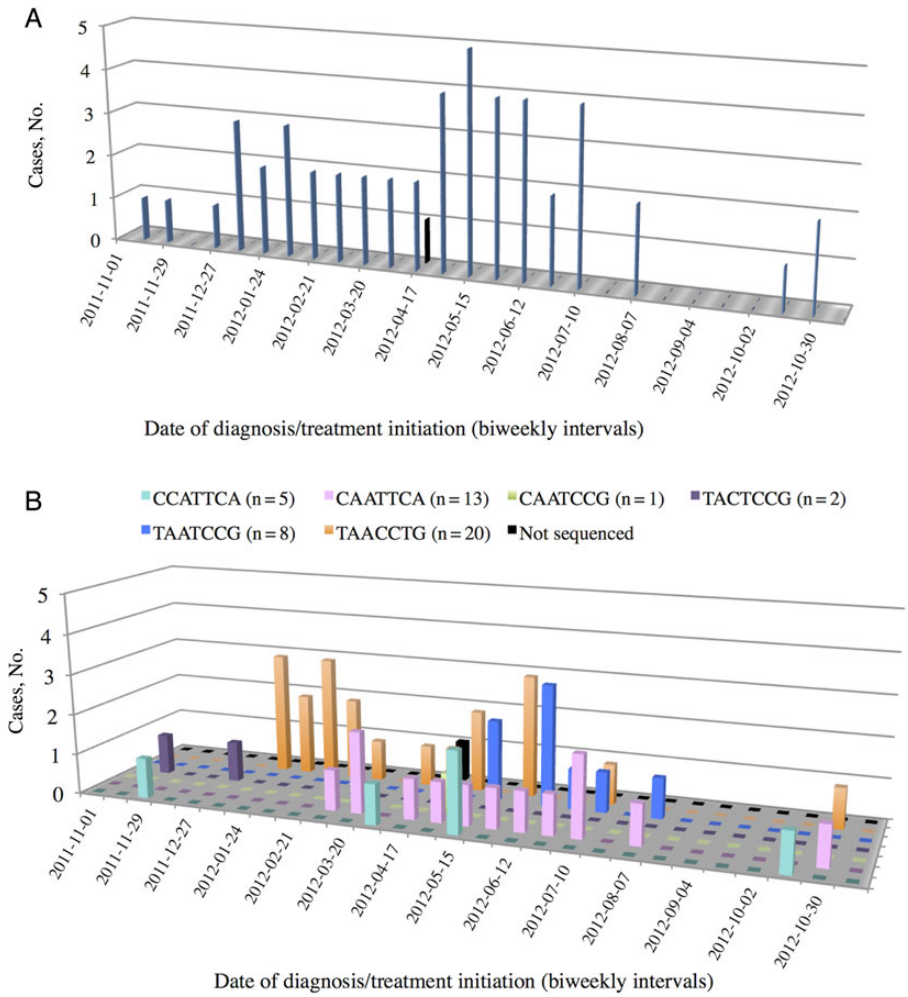


Figure 6. Epidemiologic curves of the outbreak. *A*, Overall. The numbers of cases during the outbreak are shown by date of diagnosis (year-month-date). Blue represents isolates for which whole-genome sequencing (WGS) was successful; black, isolates without WGS. There were no cases before November in 2011. *B*, Epidemiologic curve of the outbreak, stratified by WGS/epidemiologic subgroup. The numbers of cases during the outbreak are shown by date of diagnosis (year-month-date), in biweekly intervals. Cases are stratified by subgroup genotype, as indicated.

explosion of tuberculosis cases in 2011–2012. Whereas WGS alone revealed 3 clusters in the 2011–2012 outbreak, the combination of WGS with epidemiologic data allowed us to resolve this into a minimum of 6 events—5 transmission chains and 1 isolated case of reactivation. Together, these findings suggest that (1) even a single reactivation event can lead to numerous cases in this community and (2) the outbreak of 2011–2012 was not a single, rare occurrence but rather multiple smaller concurrent events. This suggests that this community is highly vulnerable to tuberculosis outbreaks, such that ongoing surveillance and vigilance against tuberculosis are warranted.

Our analysis of the outbreak leads us to several important conclusions. First, the outbreak was not due to the introduction of a new *M. tuberculosis* lineage. The isolates circulating in 2011–2012 differed by a maximum of 8 SNPs from those already present in 2007, and both IIIA and IIIB cases were documented in the years before the outbreak. Although we cannot exclude the possibility that the 2 nonsynonymous SNPs in strain IIIC affect bacterial fitness or virulence, this strain was responsible for less than half of the outbreak cases. It is therefore unlikely that these few mutations, on their own, accounted for the dramatic case rate of 2012. Rather, our findings suggest that the 2011–2012 outbreak involved the expansion of extant bacteria, consistent with a historical study of tuberculosis in Western Canada [17].

Second, both the WGS data and the clinical/epidemiologic data point to multiple transmission events, rather than a single outbreak. Although it remains possible that a single patient harbored a diversity of strains [18] and was therefore the sole source, such an explanation is neither likely nor necessary to explain the outbreak. Within a few years of the introduction of strain III, there were highly contagious carriers of each of IIIA, IIIB, and IIIC, each with epidemiologic links to multiple contacts sharing the same genotype. The knowledge that there are 3

Table 1. Household and Social Contacts With Active Tuberculosis of the Same Genotype for Each Smear-Positive Case by WGS Epidemiologic Subgroup[a]

Subgroup by WGS and Epidemiology	Date of Diagnosis		Smear Grade	Contacts With Same Genotype/Total Contacts, No. (%)[b]	
	1st Case	Smear-Positive Cases		Household Contacts	Social Contacts
IIIA (n = 1)	May 2012	. . .	. . .	0/0	0/18 (0)
IIIA (n = 2)	November 2011	November 2011	3+	1/4 (25)	0/30 (0)
IIIA (n = 8)	May 2012	June 2012	4+	0/0	3/10 (30)
		June 2012	3+	1/1 (100)	4/9 (44)
		June 2012	2+	1/1 (100)	1/3 (33)
IIIB (n = 5)	December 2011	December 2011	4+	0/0	4/32 (13)
		October 2012	3+	0/3 (0)	2/22 (9)
IIIB (n = 13)	March 2012	May 2012	2+	2/2 (100)	3/21 (14)
IIIC (n = 20)	January 2012	January 2012	3+	3/3 (100)	12/31 (39)
		April 2012	2+	1/3 (33)	5/20 (25)
		May 2012	2+	1/1 (100)	8/23 (35)
Total	. . .	. . .	. . .	10/18 (56)	42/219 (19)[c]

Abbreviation: WGS, whole-genome sequencing.

[a] Smear positive was defined as 1+ or higher, except the first subgroup comprised only 1 person, who had smear-negative disease.

[b] Because different sources named the same contacts, the denominators of contacts who developed active tuberculosis exceed the number of unique cases in the year.

[c] A 2-sample *z* test was used to assess difference in proportions ($P < .001$).

clusters (IIIA, IIIB, and IIIC) in combination with epidemiologic data has also helped identify a case of exogenous reinfection that would otherwise have been overlooked given the absence of MIRU variability. In addition, the cluster-defining SNPs of IIIA/B/C are now being used to investigate the sources of 2013–2014 cases and to distinguish relapse from reinfection in recurrent cases.

Finally, whereas MIRU and RFLP of this community would suggest that there is, and has been, ongoing transmission in this village for decades [1], WGS data challenge this interpretation. Strains I and II disappeared in 1996 and 2004, respectively, before the introduction of strain III. Given that strain III was first seen in village K and differs from strains I and II by approximately 40 SNPs, the most plausible explanation is a single reactivation case due to an organism acquired in the same village, decades before the period sampled. The majority of adults in the village have positive tuberculin skin test results, and many have chronic pulmonary diseases, so it is possible that one such individual developed transmissible disease without medical suspicion of tuberculosis, leading to the introduction of strain IIIA in 2007.

It remains unclear why this population was at such a high risk after the reappearance of tuberculosis in 2007. Given that one of the potential source cases in the outbreak presented to the clinic 4 months after symptom onset, patient delay may be a considerable factor in this population. Furthermore, although the majority of household contacts with tuberculosis shared the same genotype as the most transmissible cases within each subgroup, 44% of these household contacts did not, supporting the findings of Verver et al [19] that in an environment with high tuberculosis transmission, the traditional stone-in-pond principle may not suffice for identifying and interrupting transmission. As implemented in 1954 in Alaska [20], community-wide interventions, such as chest radiographic screening, may be needed to halt tuberculosis transmission in this setting. BCG vaccination was already reinstituted in the village in response to this outbreak after its cessation in 2005.

The primary limitation of this study is the relatively small sample size of the subgroups revealed by WGS. Despite the extraordinary incidence of disease, there was insufficient power to conduct a rigorous statistical comparison between cases in the different transmission chains. Another potential limitation is that we were unable to sequence 4 of 82 isolates. However, because we successfully sequenced 95% of all isolates from village K between 1991 and 2012, there is minimal risk of sampling bias. Finally, from a public health perspective, we were unable to identify a single, unifying cause of the 2011–2012 outbreak; this is not surprising, however, given that in-depth analysis revealed the outbreak was in fact due to ≥6 epidemiologically distinct events.

There are a number of important strengths of this study. The unique environment, with nearly all isolates sharing the same MIRU pattern, provided the opportunity to examine how limited classic genotyping methods can actually be. We have demonstrated that although isolates in a transmission chain share the same MIRU, the converse does not necessarily hold true—a fact that may have important implications for public health investigation of MIRU-defined clusters. The analysis by WGS of a single geographically isolated village provided an unexpected opportunity to witness both the disappearance and reemergence of tuberculosis over time. Isolates sequenced had a

minimum coverage of 21×, and 97% of the SNPs identified had a Phred score of >100, equivalent to a 1 in 10^{10} chance of error. The results for the outbreak obtained using the maximum likelihood method were concordant with both the previously established rate of mutation of *M. tuberculosis* [3, 5, 21, 22] and independent results from Nunavik outside village K (Supplementary Figure 3). Our phylogeny also proved robust to use of an alternate model of nucleotide substitution. We obtained independent confirmation of the 4 cluster-defining SNPs for clusters IIIA, IIIB, and IIIC using Sanger sequencing, and a previous study by Domenech et al [8] also showed very low false-positive rates using the same WGS pipeline. Finally, detailed clinical epidemiologic data were available for all cases, facilitating the verification of transmission identified by WGS.

In summary, the use of WGS permitted a fine-level analysis of an ongoing tuberculosis epidemic in this vulnerable population. The reappearance of *M. tuberculosis* was followed several years later by an epidemiologic amplification, leading to a multipronged outbreak affecting >5% of the population. Further consideration of the potential mechanisms of tuberculosis spread in this village, and other communities in Nunavik, is warranted to derive strategies to help these and other vulnerable communities control and ultimately eliminate tuberculosis.

Supplementary Data

Supplementary materials are available at The *Journal of Infectious Diseases* online (http://jid.oxfordjournals.org). Supplementary materials consist of data provided by the author that are published to benefit the reader. The posted materials are not copyedited. The contents of all supplementary data are the sole responsibility of the authors. Questions or messages regarding errors should be addressed to the author.

Notes

Acknowledgments. We thank the village council and the residents of Kangiqsualujjuaq for their collaboration and engagement in this study. We also thank the staff of Centre Local de Services Communautaires Palaqsivik for their dedicated care of patients and contacts during the outbreak. Thanks to Genevieve de Bellefeuille, BSc (Agence de la Santé et des Services Sociaux de l'Estrie), for her hard work collecting clinical and epidemiologic data during the outbreak; Isabelle Rocher, MSc (Institut National de Santé Publique du Québec), for assisting with data entry while working clinically during the outbreak; and Erwin Schurr, PhD (The Research Institute of MUHC), for his input on the genetic analysis. We also thank the NRBHSS for detailed collection of clinical and epidemiologic data for the duration of the study.

Financial support. This work was supported by the Canadian Institutes of Health Research (MOP No. 125858).

Potential conflicts of interest. All authors: No reported conflicts.

All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

References

1. Nguyen D, Proulx JF, Westley J, Thibert L, Dery S, Behr MA. Tuberculosis in the Inuit community of Quebec, Canada. Am J Resp Crit Care Med **2003**; 168:1353–7.
2. Gardy JL, Johnston JC, Ho Sui SJ, et al. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N Engl J Med **2011**; 364:730–9.
3. Walker TM, Ip CL, Harrell RH, et al. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. Lancet Infect Dis **2013**; 13:137–46.
4. Schurch AC, Kremer K, Daviena O, et al. High-resolution typing by integration of genome sequencing data in a large tuberculosis cluster. J Clin Microbiol **2010**; 48:3403–6.
5. Roetzer A, Diel R, Kohl TA, et al. Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiological study. PLoS Med **2013**; 10:e1001387.
6. Kato-Maeda M, Ho C, Passarelli B, et al. Use of whole genome sequencing to determine the microevolution of *Mycobacterium tuberculosis* during an outbreak. PLoS One **2013**; 8:e58235.
7. van Soolingen D, Hermans PW, de Haas PE, Soll DR, van Embden JD. 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. J Clin Microbiol **1991**; 29:2578–86.
8. Domenech P, Rog A, Moolji JUD, et al. The origins of a 350-kilobase genomic duplication in *Mycobacterium tuberculosis* and its impact on virulence. Infect Immun **2014**; 82:2902–12.
9. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol **2011**; 28:2731–9.
10. Nei M, Kumar S. Molecular evolution and phylogenetics. New York, NY: Oxford University Press, **2000**.
11. Tamura K. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. Mol Biol Evol **1992**; 9:678–87.
12. Jukes TH, Cantor CR. Evolution of protein molecules. New York, NY: Academic Press, **1969**:21–132.
13. Centers for Disease Control. Guidelines for the investigation of contacts of persons with infectious tuberculosis. MMWR Recommend Rep **2005**; 54:1–62.
14. Statistics Canada. 2012. Kangiqsualujjuaq, Quebec (Code 2499090) and Quebec (Code 24) (table). Census Profile. 2011 Census. Statistics Canada Catalogue no. 98-316-XWE, Ottawa. Released October 24, 2012. http://www12.statcan.gc.ca/census-recensement/2011/dp-pd/prof/index.cfm?Lang=E. Accessed 10 December 2014.
15. Felsenstein J. Confidence-limits on phylogenies: an approach using the bootstrap. Evolution **1985**; 39:783–91.
16. Comas I, Coscolla M, Luo T, et al. Out-of-Africa migration and neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. Nat Genet **2013**; 45:1176–82.
17. Pepperell CS, Granka JM, Alexander DC, et al. Dispersal of *Mycobacterium tuberculosis* via the Canadian fur trade. Proc Natl Acad Sci USA **2011**; 108:6526–31.
18. Perez-Lago L, Comas I, Navarro Y, et al. Whole genome sequencing analysis of intrapatient microevolution in *Mycobacterium tuberculosis*: potential impact on the inference of tuberculosis transmission. J Infect Dis **2013**; 209:98–108.
19. Verver S, Warren RM, Munch Z, et al. Proportion of tuberculosis transmission that takes place in households in a high-incidence area. Lancet **2004**; 363:212–4.
20. Grzybowski S, Styblo K, Dorken E. Tuberculosis in Eskimos. Tubercle **1976**; 57(suppl 4):S1–58.
21. Ford CB, Shah RR, Maeda MK, et al. *Mycobacterium tuberculosis* mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. Nat Genet **2013**; 45:784–90.
22. Bryant JM, Schürch AC, van Deutekom H, et al. Inferring patient to patient transmission of *Mycobacterium tuberculosis* from whole genome sequencing data. BMC Infect Dis **2013**; 13:110.

APPENDIX 2-2

Supplementary Data:

Lee RS, Radomski N, Proulx J-F, Manry J, McIntosh F, Desjardins F, Soualhine H, Domenech P, Reed MB, Menzies D, Behr MA. Reemergence and re-amplification of tuberculosis in the Canadian Arctic. *J Infect Dis* 2015;211(12):1905-1914

Re-emergence and Amplification of Tuberculosis in the Canadian Arctic

Robyn S. Lee, BSc[1–3†], Nicolas Radomski, PhD[3,†], Jean-Francois Proulx, MD[4], Jeremy Manry, PhD[2,3,5], Fiona McIntosh, BSc[3], Francine Desjardins, DTL[6], Hafid Soualhine, PhD[7], Pilar Domenech, PhD[3], Michael B. Reed, PhD[2,3], Dick Menzies, MD[3,8], Marcel A. Behr, MD[2,3*]

Supplementary Data

Detailed Methods

Genomics. MTB isolates were cultured once on Middlebrook 7H10 agar and then genomic DNA (gDNA) extractions were performed as per van Soolingen et al (1).

gDNA fragments were multiplexed by 24 for Paired-end 250 bp sequencing using the MiSeq 250 System (Illumina). Reads were then trimmed using Trimmomatic (v.0.25, (2)) to retain base-pairs (bp) with a Phred33 score ≥30 (corresponding to 99.9% accuracy, where Phred is -$10*\log_{10}P_{error}$) and a minimum read-length of 50 bp. Reads were deposited in the National Center for Biotechnology Information's Sequence Read Archive, under Accession number SRP039605. Reads were aligned to the H37Rv reference genome (NCBI Accession number NC_000962.3) using the Burrows-Wheeler Aligner (v.0.6.2 and v.7.1.0, (3)). Variants were called using a Bayesian genotype likelihood model (Unified Genotyper, Genome Analysis Toolkit, Broad Institute, v.2.7.4) and only SNPs with a Phred score >50 were retained for analysis. SNPs were annotated with the uid57775 database for the H37Rv reference genome using SnpEff (v.3.3h, (4)) and confirmed using Artemis (v.15.0.0, Sanger Institute).

Assessing for contamination:

Pairs of MTB isolates that had 0 SNPs difference were investigated for potential cross-contamination by determining whether they were processed in the same batch of samples.

Ascertainment bias:

Ascertainment bias was assessed in two manners: 1) by comparing Phred scores of SNPs across genomes to verify that Phred scores were consistently high, regardless of depth of coverage, and 2) by confirming cluster-defining SNPs (in genes *Rv0828c*, *carB*, *Rv1835c*, and *Rv3263*) with Sanger sequencing for isolates (6 isolates for each cluster IIIA, IIIB and IIIC) presenting the lowest values of coverage.

Mixed infection: We assessed for mixed infection by verifying that the frequency of heterogeneous alleles was constant across each genome and examining the corresponding Phred scores of these alleles.

Mycobacterial interspersed repetitive units (MIRU) and IS*6110* restriction fragment length polymorphism (RFLP): MIRU data for the outbreak year were provided by the Laboratoire de Santé Publique du Québec. IS*6110* RFLP for the outbreak isolates were conducted by Southern blotting (5), with analysis as described in (6).

Figure Legends

Supplementary Figure 1. Classical Molecular Typing of Village K isolates.

Legend. 1A. Mycobacterial Interspersed Repetitive Units (MIRU). 24 loci MIRU shown for all 50 cases in the 2011-2012 outbreak. Pink – 47/50 isolates had identical MIRU patterns. Purple – These 2 cases had epidemiologic links with isolates with MIRU 224325143324234534423463. One MIRU locus missing, identical to main cluster at 23/24 sites. Blue – This isolate is 1 locus different from main cluster. WGS was unsuccessful for this isolate. **1B. IS*6110* Restriction Fragment Length Polymorphism (RFLP).** – RFLP shown for representative isolates from all 3 clusters of strain III (A/B/C), 2011-2012; all shared an identical pattern. Two isolates from each of strains I and II shown for comparison.

Supplementary Figure 2. Distribution of Single Nucleotide Polymorphisms (SNPs) and Heterogeneous Alleles per MTB Genome according to Phred Score Ranges

Legend. On the left: The distribution of SNPs for each genome. On the right: the number of heterogeneous alleles for each genome is indicated. The Phred score ranges of these SNPs are indicated at the top. Genomes are sorted from lowest highest depth of coverage (X in brackets). The number of SNPs per genome was constant at 1,063 (SD ± 32) with Phred score >50 and the Phred scores of the majority of SNPs were high (between 500 to 5000) for genomes with low or high depth of coverage, indicating the majority of SNPs are correctly called. The number of heterogeneous alleles per genome was constant at 324 (SD ± 82) with Phred scores >50, indicating that no multiple strain infections were present. The Phred scores of the majority of the heterogeneous alleles are low (between 50 to 500).

Supplementary Figure 3. Single Nucleotide Polymorphisms (SNPs) Associated with 'Improbable' Transmission Events in Other Villages of Nunavik

Legend. Contact investigation data were obtained for all microbiologically confirmed TB cases diagnosed in the other villages of Nunavik (excluding village K), years 2006-2012. Epidemiologic connections between cases were classified by a single clinician. Case pairs were designated as 'Improbable transmission' when cases with no known epi links resided in different villages. WGS was conducted on 42 isolates from other villages for which these epidemiologic data were available and a pairwise SNP matrix was generated. Pairwise comparisons between 'Improbable' transmission' pairs were included in the final analysis, yielding a total of 631 pairwise comparisons. The median pairwise SNP distance was 53 (IQR 13-56) between 'Improbable transmission' pairs. No two cases designated 'Improbable transmission' were within 0-1 SNPs of one another. 34 pairwise comparisons associated with one isolate were excluded from the above figure, as this isolate was >770 SNPs from all others.

References

1. van Soolingen D, Hermans PW, de Haas PE, Soll DR, van Embden JD. 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. J Clin Microbiol **1991**; 29(11):2578–2586.

2. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **2014**; 30(15):2114–2120.

3. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics **2010**; 26(5):589–595.

4. Cingolani P, Platts A, Wang LL, 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 **2012**; 6(2):80–92.

5. van Embden JD, Cave MD, Crawford JT, et al. Strain identification of Mycobacterium tuberculosis by DNA fingerprinting: recommendations for a standardized methodology. J Clin Microbiol **1993**; 31(2):406–409.

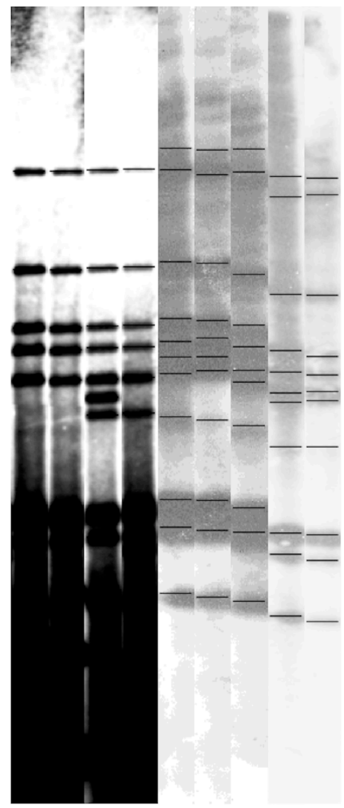
6. Nguyen D, Proulx J-F, Westley J, Thibert L, Dery S, Behr MA. Tuberculosis in the Inuit Community of Quebec, Canada. Amer J Resp Crit Care Med **2003**; 168(11):1353–1357.

Supplementary Figure 1

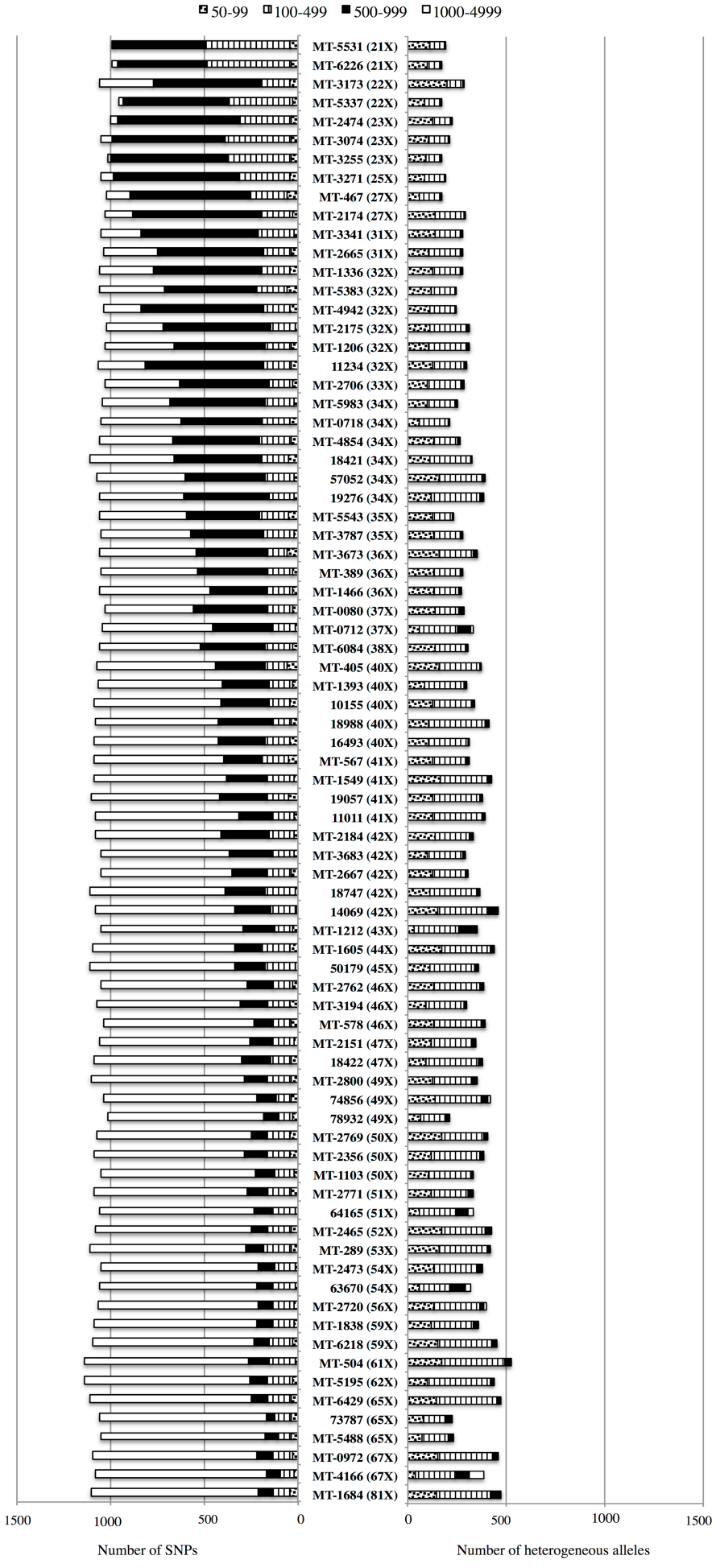
A

22432514332423453442 463
224325143324234534423 63
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325143324234534423463
224325142324234534423463

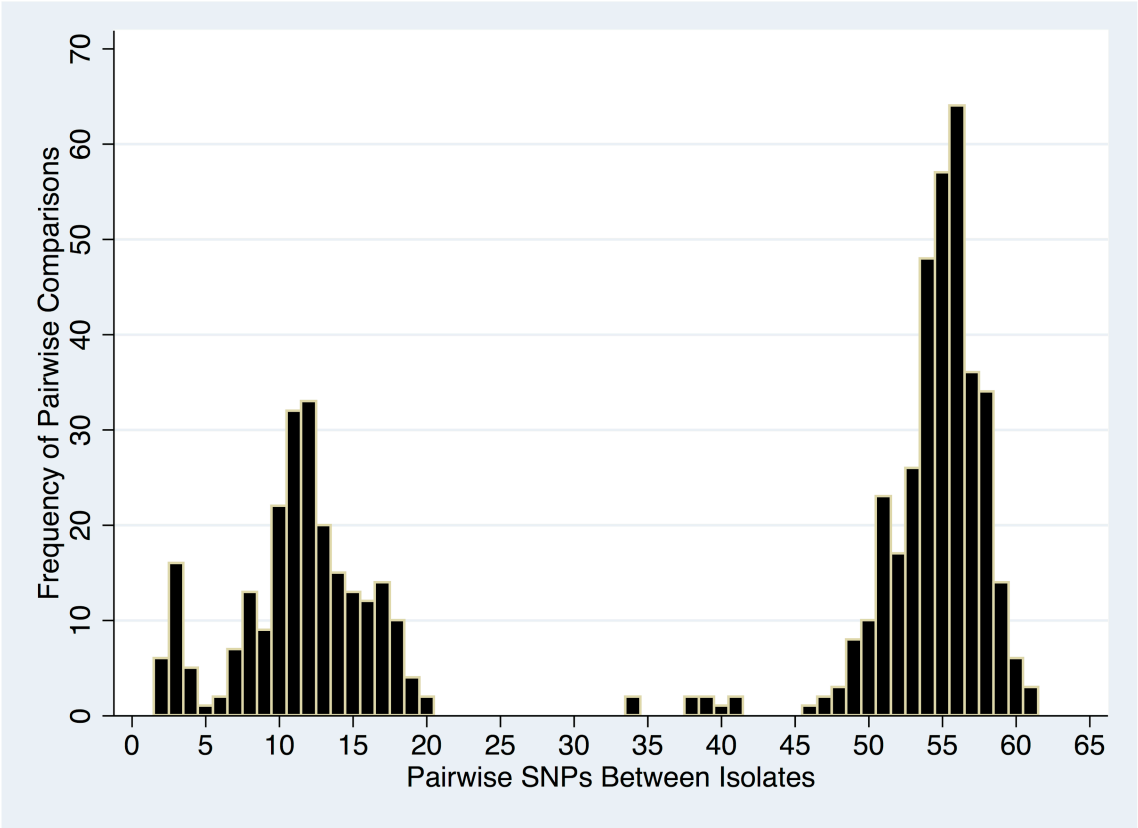
B



Supplementary Figure 2



Supplementary Figure 3



APPENDIX 3-1

Reprint of:

Lee RS, Radomski N, Proulx J-F, Levade I, Shapiro BJ, McIntosh F, Soualhine H, Menzies D, Behr MA. Population genomics of *Mycobacterium tuberculosis* in the Inuit. *Proc Natl Acad Sci USA* 2015;112(44):13609-13614

Population genomics of *Mycobacterium tuberculosis* in the Inuit

Robyn S. Lee[a,b,c,1], Nicolas Radomski[b,c,1], Jean-Francois Proulx[d], Ines Levade[e], B. Jesse Shapiro[e], Fiona McIntosh[b,c], Hafid Soualhine[f], Dick Menzies[b,c,g], and Marcel A. Behr[b,c,2]

[a]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, QC, Canada, H3A 1A2; [b]The Research Institute of the McGill University Health Centre, Montreal, QC, Canada, H4A 3J1; [c]McGill International TB Centre, McGill University Health Centre, Montreal, QC, Canada, H4A 3J1; [d]Nunavik Regional Board of Health and Social Services, Kuujjuaq, QC, Canada, J0M 1C0; [e]Département de Sciences Biologiques, Université de Montréal, Montreal, QC, Canada, H2V 2S9; [f]Laboratoire de Santé Publique du Québec, Sainte-Anne-de-Bellevue, QC, Canada, H9X 3R5; and [g]Respiratory Epidemiology and Clinical Research Unit, Montreal Chest Institute, Montreal, QC, Canada, H4A 3J1

Edited by Carl F. Nathan, Weill Cornell Medical College, New York, NY, and approved September 16, 2015 (received for review April 13, 2015)

Nunavik, Québec suffers from epidemic tuberculosis (TB), with an incidence 50-fold higher than the Canadian average. Molecular studies in this region have documented limited bacterial genetic diversity among *Mycobacterium tuberculosis* isolates, consistent with a founder strain and/or ongoing spread. We have used whole-genome sequencing on 163 *M. tuberculosis* isolates from 11 geographically isolated villages to provide a high-resolution portrait of bacterial genetic diversity in this setting. All isolates were lineage 4 (Euro-American), with two sublineages present (major, $n = 153$; minor, $n = 10$). Among major sublineage isolates, there was a median of 46 pairwise single-nucleotide polymorphisms (SNPs), and the most recent common ancestor (MRCA) was in the early 20th century. Pairs of isolates within a village had significantly fewer SNPs than pairs from different villages (median: 6 vs. 47, $P < 0.00005$), indicating that most transmission occurs within villages. There was an excess of nonsynonymous SNPs after the diversification of *M. tuberculosis* within Nunavik: The ratio of nonsynonymous to synonymous substitution rates (dN/dS) was 0.534 before the MRCA but 0.777 subsequently ($P = 0.010$). Nonsynonymous SNPs were detected across all gene categories, arguing against positive selection and toward genetic drift with relaxation of purifying selection. Supporting the latter possibility, 28 genes were partially or completely deleted since the MRCA, including genes previously reported to be essential for *M. tuberculosis* growth. Our findings indicate that the epidemiologic success of *M. tuberculosis* in this region is more likely due to an environment conducive to TB transmission than a particularly well-adapted strain.

Mycobacterium tuberculosis | evolution | whole-genome sequencing

The tubercule bacillus, *Mycobacterium tuberculosis*, is a highly successful, medically important human-adapted pathogen. Studies of diverse strain collections reveal a geographic aggregation of the principal *M. tuberculosis* lineages (1) consistent with a dissemination of this organism around the world with the paleo migration (2). Ancient DNA studies also support the notion that *M. tuberculosis* has caused disease in humans for thousands of years. Thus, it can be inferred that *M. tuberculosis* has evolved in step with its human host, successfully responding to changes in the host and its environment that could affect the capacity to cause transmissible disease.

In contrast to the global diversity of *M. tuberculosis* strains (1–3), we have previously observed limited genetic diversity in the Nunavik region of Québec (4). One possible explanation is a founder strain, wherein genetic similarity is due to a single recent introduction of a bacterium and may not necessarily represent ongoing spread between communities. In this scenario, isolates might have indistinguishable genotypes by conventional genotyping modalities (restriction fragment length polymorphism, mycobacterial interspersed repetitive units, spoligotyping) but distinct genotypes when assessed using a higher-resolution method, namely whole-genome sequencing (WGS) (5). An additional explanation is that a single clone of *M. tuberculosis* is currently spreading both within and between villages; however, the great distances between these communities that are not linked by roads make intervillage spread less likely. These possible explanations need not be mutually exclusive.

To evaluate these possibilities, we conducted WGS on *M. tuberculosis* isolates from Nunavik isolated over 23 y. Estimation of the divergence date of the most recent common ancestor (MRCA) provided evidence that tuberculosis (TB) was introduced into this region in the early 20th century, following which time there has been substantial ongoing transmission, predominantly within villages. This setting provides a unique opportunity to study the genomic characteristics of an epidemiologically successful strain of *M. tuberculosis* over time.

Results

Whole-Genome Sequencing and Lineage Identification. There were 149 microbiologically confirmed TB cases diagnosed in Nunavik between 2001 and 2013; we obtained high-quality WGS data for 137/149 (92%). An additional 26 genomes were successfully sequenced from strains previously sampled between 1990 and 2000 (4). In total, WGS was conducted on 163 *M. tuberculosis* isolates. The average depth of coverage was 44.6× across 99.6% of the H37Rv reference genome.

All 163 genomes from the Nunavik region presented the 7-bp deletion in polyketide synthase (*pks*) *15/1* that characterizes lineage 4 of *M. tuberculosis* (the Euro-American lineage) (6). By comparing

Significance

Through an in-depth analysis of whole-genome sequencing data from Nunavik, Québec, we inferred the evolution of a single dominant strain of *Mycobacterium tuberculosis*. Our analyses suggest that *M. tuberculosis* was first introduced into this region in the early 20th century. Since this time, *M. tuberculosis* has spread extensively, predominantly within but also between villages. Despite a genomic profile that lacks features of a hypervirulent strain, this strain has thrived in this region and continues to cause outbreaks. This suggests that successful clones of *M. tuberculosis* need not be inherently exceptional; host or social factors conducive to transmission may contribute to the ongoing tuberculosis epidemic in this and other high-incidence settings.

Author contributions: J.-F.P., B.J.S., D.M., and M.A.B. designed research; R.S.L., N.R., F.M., and H.S. performed research; J.-F.P., I.L., B.J.S., F.M., and H.S. contributed new reagents/analytic tools; R.S.L., N.R., I.L., B.J.S., and M.A.B. analyzed data; and R.S.L., N.R., D.M., and M.A.B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: The aligned reads reported in this paper have been deposited in the National Center for Biotechnology Information's Sequence Read Archive (accession no. SRP039605, BioProject PRJNA240330).

[1]R.S.L. and N.R. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: marcel.behr@mcgill.ca.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1507071112/-/DCSupplemental.

the Nunavik isolates with three genomes from each of the *M. tuberculosis* lineages (1–7), we observed that the 163 genomes were tightly clustered in two distinct sublineages: one consisting of 153 isolates (major; Mj) and the other consisting of 10 isolates (minor; Mn) (Fig. 1). Phylogenetic analyses based on single-nucleotide polymorphisms (SNPs) (Figs. 1 and 2) were supported by deletions confirmed by PCR (Fig. S1 and Dataset S1).

Excluding SNPs in PE/PGRS and PPE genes, as well as mobile elements, as these may be at higher risk of false positives (5, 7), 1,288 single-nucleotide polymorphic loci were included comparing all genomes together against H37Rv (Dataset S2). The 153 isolates of the Mj sublineage had an average of 674 SNPs compared with H37Rv; the 10 isolates of the Mn sublineage had an average of 451 SNPs. There were 442 SNP loci shared across all Mj isolates, unique to this sublineage, and 214 SNP loci shared by all 10 Mn isolates that were not present in the Mj sublineage. According to the barcode proposed by Coll et al. (8) and the PhyTb tool of the PhyloTrack library (pathogenseq.lshtm.ac.uk/phytblive/index.php), the Mj and Mn sublineages can be classified as *M. tuberculosis* 4.1.2 and 4.8, respectively.

Phylogenetic analysis and the geographic distribution of isolates further distinguished the Mj and Mn sublineages (Fig. 2). To quantify diversity, we determined the number of pairwise SNPs within each sublineage. Among isolates of the Mj sublineage, the median number of pairwise SNPs was 46 [interquartile range (IQR) 13–49], with a maximum of 72. For isolates of the Mn sublineage, the median number of pairwise SNPs was 1 (IQR 0–2), with a maximum of 22. Nine of the 10 isolates from this sublineage were from the same village.

Transmission Occurs Mostly Within Villages. To evaluate where most ongoing transmission occurs, we examined pairwise SNPs between isolates of the Mj sublineage, within and between villages, as these comprised over 90% of the cases in this region. The median number of pairwise SNPs was significantly lower for intravillage pairs (6, IQR 3–46) than for intervillage pairs (47, IQR 44–50, Wilcoxon–Mann–Whitney, $P < 0.00005$). For both intra- and intervillage comparisons, a bimodal distribution was evident (Fig. 3). For the intravillage pairs ($n = 3,689$), the first mode comprised 61% of all pairwise comparisons and had a median of 3 SNPs (IQR 2–5). For the intervillage pairs ($n = 7,939$), the first mode comprised only 12% of all pairwise comparisons, and had a median of 9 SNPs (IQR 6–13). For both intra- and intervillage pairs, the second mode had similar distributions (median 47, IQR 45–49 and median 48, IQR 45–50, respectively), consistent with the star-like pattern shown in Figs. 1 and 2.

We also considered thresholds for transmission based on published *M. tuberculosis* substitution rates (0.5 SNPs per genome per y, 95% confidence interval 0.3–0.7) (5, 9). For a study spanning 23 y (1991–2013 inclusive), we expected that epidemiologically linked cases would be separated by no more than 12 SNPs. Applying this threshold, 2,208 of 3,689 (60%) intravillage pairs were separated by 12 or fewer SNPs, compared with 683 of 7,939 (9%) intervillage pairs (two-sample z test for difference in proportions, $P < 0.00005$). Sensitivity analyses applying substitution rates of 0.3 and 0.7 SNPs per genome per y yielded similar results.

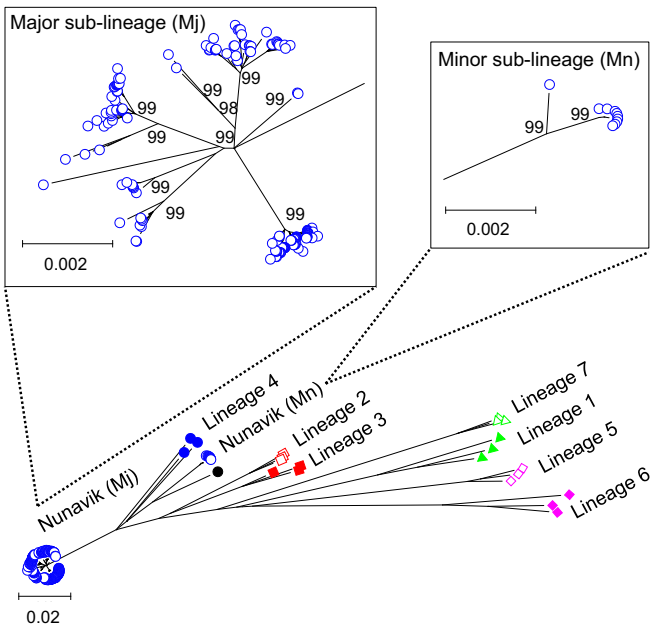


Fig. 1. Maximum likelihood tree of 163 *M. tuberculosis* isolates from Nunavik and 21 representative genomes of lineages 1–7. Phylogenetic clusters based on 9,016 single-nucleotide polymorphic loci identified across 184 genomes compared with H37Rv (solid black circle). The scale bars represent the number of substitutions per site. Bootstrap values from 1,000 replicates are shown for branches within the Mj and Mn sublineages. For clarity, only values ≥98 are shown.

***M. tuberculosis* Diversified in Nunavik During the 20th Century.** Relative to the global genetic diversity of *M. tuberculosis*, the total diversity of strains in Nunavik was low, consistent with a recent introduction of TB into this region. To evaluate this hypothesis, we estimated the MRCAs for each sublineage using Bayesian molecular dating (10, 11). Constraining the substitution rate of *M. tuberculosis* based on previous estimates (5, 9) we inferred the MRCA of the Mj sublineage to be 1919 [95% highest posterior density interval (HPD) 1892–1946], with other divergence dates within the Mj sublineage scattered over the 20th century (Table 1, analysis 1). The Mn sublineage was found to have an MRCA of 1976 (95% HPD 1951–1994). Repeating these analyses without constraining the substitution rate yielded similar results (Table 1, analyses 2 and 3).

Natural Selection of *M. tuberculosis* in a New Environment. The *M. tuberculosis* population may have experienced a new regime of natural selection upon its introduction into Nunavik. First, *M. tuberculosis* could have experienced a population bottleneck upon introduction, reducing the efficacy of natural selection and allowing the fixation of deleterious mutations. Second, upon entry into a new environment, *M. tuberculosis* could have experienced positive selection, retaining fitter variants over time. Third, if the environment was conducive to transmission of *M. tuberculosis*, there may have been a relaxation of purifying selection across the entire genome. These scenarios are not mutually exclusive, and other scenarios are possible as well.

To measure natural selection at the protein level, we used the ratio of nonsynonymous to synonymous substitution rates (dN/dS), reasoning that this should remain stable over time in the absence of changing regimes of natural selection (12). Specifically, we tested the null hypothesis that dN/dS remained the same pre- and post-diversification of each *M. tuberculosis* sublineage. We first reconstructed the ancestral sequences of the MRCA for each sublineage, along with that of the common ancestor for these two sublineages (denoted "Mj–Mn"). We then compared the nonsynonymous and synonymous SNPs (nsSNPs and sSNPs, respectively) between these reconstructed ancestors (Mj–Mn versus Mj, Mj–Mn versus Mn) to obtain the dN/dS for each sublineage prediversification. To calculate the dN/dS postdiversification (i.e., subsequent to the MRCAs for each sublineage), we generated a concatenated sequence of codons for both the Mj and Mn sublineages that included all SNP loci and compared each sequence with that of its respective ancestor. In this phylogenetic approach, each independent SNP was counted exactly once (i.e., SNPs present in multiple isolates were not recounted). In total, for the Mj sublineage, we identified 229 nsSNPs and 154 sSNPs before its introduction into Nunavik, compared with 238 nsSNPs and 107 sSNPs that occurred subsequently (Dataset S2). The dN/dS ratio for SNPs prediversification was 0.534, consistent with published estimates for *M. tuberculosis*

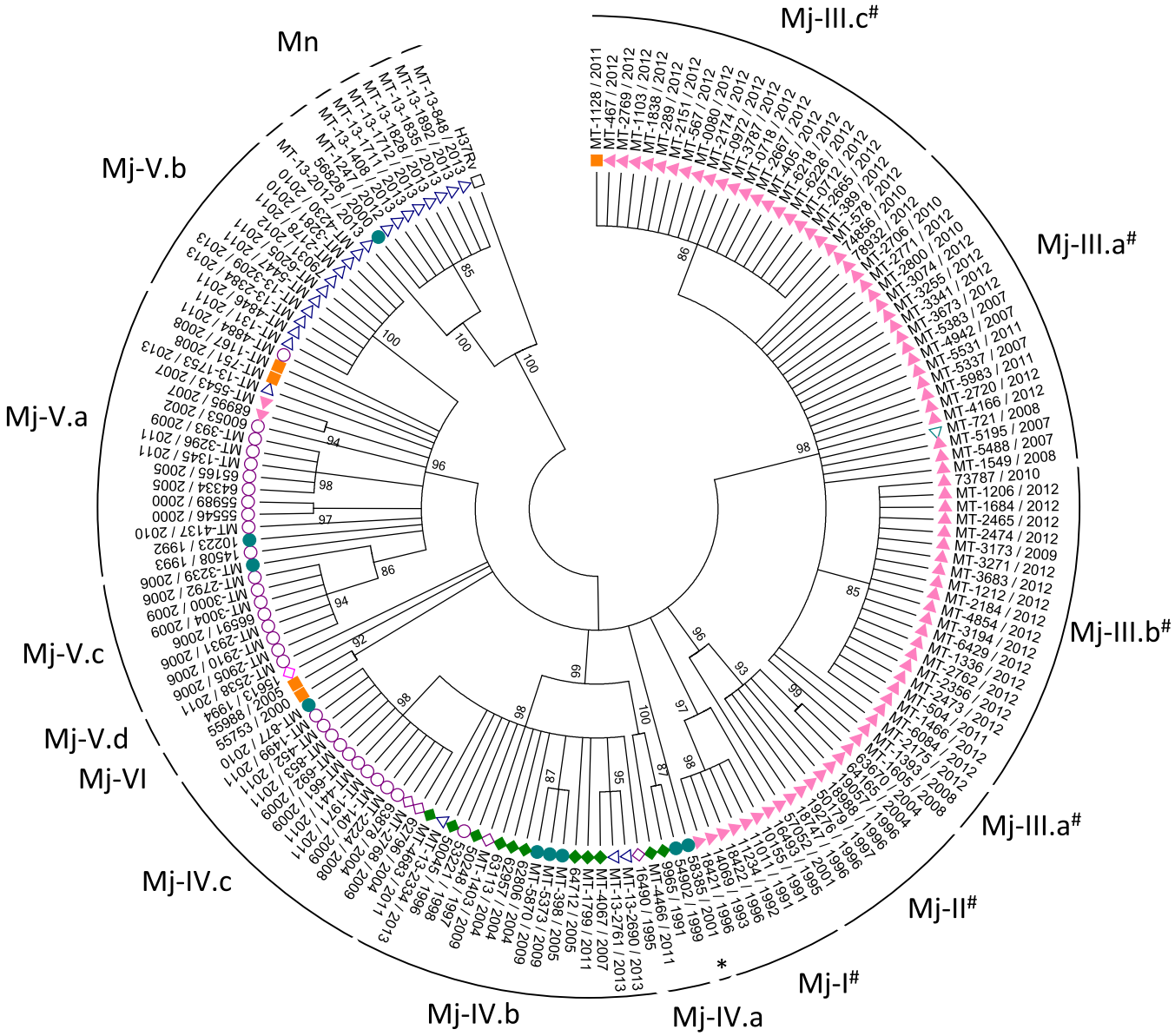


Fig. 2. Maximum likelihood tree of 163 *M. tuberculosis* isolates from Nunavik. Phylogenetic clusters were identified based on 1,288 single-nucleotide polymorphic loci compared with H37Rv. Solid and dashed lines indicate isolates of the Mj and Mn sublineages, respectively. Colored shapes represent the reference genome (bordered black square) and the villages of Nunavik: A (bordered blue triangle), B (full orange square), C (bordered purple circle), D (full green diamond), E (bordered purple diamond), K (full pink triangle), and other (full green circle). *Genome with a unique single-nucleotide polymorphism profile. #Phylogenetic clusters defined previously in ref. 32. Years of diagnosis are indicated. Bootstrap support from 1,000 replicates is shown. Branches supported by less than 80% of bootstrap replicates are collapsed.

(13), whereas the dN/dS postdiversification was 0.777 (Table 2, analysis 1a; *G* test based on numbers of nsSNPs and sSNPs pre- and postdiversification, $P = 0.010$). Singleton SNPs, present in only one isolate, are expected to be enriched in nonsynonymous mutations destined to be purged by purifying selection. To evaluate whether the increased dN/dS was attributable to these transient mutations, we restricted our analysis to SNPs present in ≥2 isolates. We still observed a significant increase in the dN/dS, going from 0.534 prediversification to 0.928 postdiversification (Table 2, analysis 1b). There was no significant difference in postdiversification nsSNPs and sSNPs comparing analyses with and without singletons (Fisher's exact test, $P = 0.472$). As an alternative method of calculating the dN/dS postdiversification, we conducted a pairwise analysis wherein the median dN/dS was obtained by comparing each of the 153 Mj isolates with its respective ancestral sequence. This yielded similar results, whether singletons were included or excluded (Table 2, analysis 2). Compared with the Mj sublineage, the dN/dS ratios for the Mn sublineage were more stable over time (Table 2).

The efficiency of purifying selection to remove deleterious nonsynonymous mutations is reduced when populations undergo dramatic size fluctuations due, for example, to bottlenecks or exponential growth. To investigate whether the increased dN/dS ratio in the Mj sublineage was due to an expanding bacterial population size over time, we constructed Bayesian skyline plots (Fig. S2). Model comparison using Akaike's information criterion for Markov chain Monte Carlo samples [AICM (14)] rejected an exponential population growth in favor of a constant population size or Bayesian skyline model (Table S1). Together, these results suggest that the genome-wide increase in dN/dS was not due to a population bottleneck followed by exponential growth, nor to a lack of time for purifying selection to purge deleterious nsSNPs.

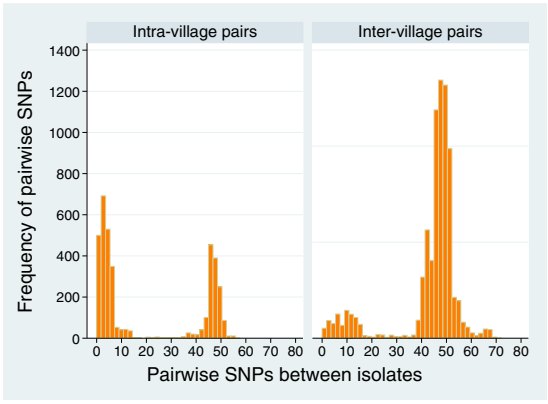


Fig. 3. Pairwise SNPs between isolates of the major sublineage of Nunavik. There were a total of 11,628 pairwise comparisons: 3,689 intravillage case pairs and 7,939 intervillage case pairs.

Genes Affected by SNPs and/or Deletions. Unlike genome-wide relaxation, wherein the whole genome is affected, positive selection is thought to target specific genes (15, 16). Across the 153 genomes of the Mj sublineage, we identified 218 and 227 genes with nsSNPs pre- and postdiversification, respectively (Dataset S2). To evaluate whether any particular categories of *M. tuberculosis* genes were unusually variable postdiversification, we tabulated these SNPs according to gene categories described in the literature (Fig. 4 and Datasets S3 and S4). There was no statistically significant difference between the proportion of genes with nsSNPs in any categories pre- versus postdiversification (two-sample z test for difference in proportions, $P > 0.05$). However, genes predicted to be conditionally essential for *M. tuberculosis* survival in vitro, in macrophages, or in vivo were not spared nsSNPs (Dataset S5). Mutations in essential genes often affected a residue that is conserved in the closely related mycobacterial species *Mycobacterium canettii* (17) and *Mycobacterium kansasii* (18), with three genes (*Rv0338c*, *echA5*, and *murC*) incurring distinct nsSNPs in different strains (Dataset S5).

In addition to these potentially deleterious SNPs, all Mj isolates lacked eight regions, resulting in 13 deleted genes. Certain strains also suffered a further seven deletions, disrupting 28 genes (Dataset S1). Certain gene categories appeared overrepresented in postdiversification deletions (e.g., genes acquired through lateral gene transfer, mobile elements), but the low number of deleted genes precluded robust statistical analysis (Fig. 4 and Dataset S1). Four genes predicted to be essential in genomic screens were completely (*Rv2335*) or partially (*Rv1939*, *Rv2885c*, and *Rv3135*) deleted in some isolates of the Mj sublineage (Dataset S3). *Rv2335* (i.e., *cysE*) codes for a serine acetyltransferase, predicted to be essential for survival in vivo (19), that was absent in eight isolates. *Rv2885c* codes for a transposase in the IS*1539* insertion sequence that is predicted to be essential for survival in vivo (19), whereas *Rv3135* codes for *PPE50* and is predicted to be essential for survival in vitro (20). *Rv1939* codes for an oxydoreductase predicted to be essential for survival in vivo (19) that was deleted in one isolate (18421) (Dataset S1).

Discussion

The Inuit originally came from eastern Siberia, via the Bering Strait, in two waves over several thousands of years (21). Given the recognized close association between *M. tuberculosis* and human populations, it is theoretically possible that they brought an East Asian lineage of *M. tuberculosis* with them to the Canadian Arctic. Our data refute this scenario by revealing only lineage 4 (Euro-American) isolates. The low amount of genetic diversity among isolates from different villages indicates that the vast majority of TB cases in this region are the consequence of a single introduction of *M. tuberculosis*, perhaps from Europe, around the early 20th century. The introduction and diversification of a single dominant clone in Nunavik provide an unobstructed view of *M. tuberculosis* over time, enabling us to draw certain inferences about the epidemiology and evolution of this highly successful human-adapted pathogen.

The Inuit have had casual interactions with Europeans since the 17th century, most notably with whalers and explorers who sailed along the coasts of Hudson's Bay and Labrador (22). However, the first permanent settlements of the Hudson's Bay Company in the region now known as Nunavik date to the late 19th and early 20th centuries, following which there were more sustained interactions between the Inuit and traders (23). Our MRCA estimates support an introduction of TB into this region during this period, which is consistent with some, but not all, historical accounts of when TB was first observed (24). The apparent lack of TB before the early 20th century, despite several centuries of Inuit–European interactions, supports that TB is generally not spread through casual contact, as is the case for measles or chickenpox. This is also consistent with our analysis of the pairwise SNPs between isolates across villages; only a small proportion of intervillage case pairs had low SNP differences, arguing against transmission during casual contact, as can occur at cultural gatherings that bring together members of different villages. Supporting this, villages often had one predominant strain, and individual strains were mostly confined to one village (Fig. 2). This observation presents both an opportunity and a challenge for public health; whereas TB should in theory be amenable to control through scaled-up efforts, it may be that village-by-village, rather than regional, interventions will be needed to interrupt transmission in this setting.

In a number of high-incidence countries, the emergence of an epidemiologically successful strain has been attributed to virulence features encoded in the bacterial genome (25). For instance, the polyketide synthase-derived phenolic glycolipid (PGL) coded by the

Table 1. Estimated year of divergence of *M. tuberculosis* sublineages and clusters of Nunavik

Phylogenetic sublineages and clusters	Analysis 1*	Analysis 2	Analysis 3
Mj–Mn	1053 (602–1450)	1243 (836–1575)	744 (230–1216)
Mj	1919 (1892–1946)	1922 (1890–1950)	1904 (1873–1930)
Mj-I-II†	1942 (1919–1964)	1947 (1921–1967)	1925 (1898–1948)
Mj-V	1952 (1929–1973)	1956 (1932–1978)	1935 (1909–1958)
Mj-IV	1965 (1949–1978)	1966 (1951–1980)	1958 (1941–1973)
Mj-III.a.b.c†	1999 (1993–2004)	1999 (1993–2004)	2000 (1993–2004)
Mj-VI	1999 (1995–2000)	1999 (1995–2000)	1999 (1995–2000)
Mn	1976 (1951–1994)	1979 (1953–1997)	1969 (1943–1987)

All numbers are expressed in calendar years, rounded to the nearest whole number. Analysis 1: calibration point, concatenated alleles. Analysis 2: no calibration point, concatenated alleles. Analysis 3: no calibration point, weighting for constant sites. The median date of divergence is shown in years, with corresponding 95% highest posterior density intervals.

*Results of this analysis are reported in the text.

†Strain code per Lee et al. (32).

Table 2. dN/dS of *M. tuberculosis* sublineages pre- and postdiversification in Nunavik

	Mj sublineage			Mn sublineage		
Analysis	Prediversification	Postdiversification	*P* value	Prediversification	Postdiversification	*P* value
1a: all SNPs	0.534	0.777	0.010	0.547	0.615	0.873
1b: excluding singletons	0.534	0.928	0.005	0.547	0.759	0.767*
2a: all SNPs	0.534	0.947	<0.00005	0.547	0.759	0.006
2b: excluding singletons	0.534	0.953	<0.00005	0.547	0.759	0.006

Ancestral sequences were reconstructed for the MRCA of the Mj–Mn sublineages, as well as the Mj sublineage and the Mn sublineage. Prediversification: 229 nonsynonymous SNPs and 154 synonymous SNPs identified in the Mj sublineage, and 113 nsSNPs and 75 sSNPs in the Mn sublineage. Analysis 1a: dN/dS prediversification was calculated by comparing ancestral sequences. For postdiversification, concatenated sequences of codons for each sublineage were generated based on all SNP loci identified, with SNPs in more than one isolate only contributing once. Overall, there were 238 nsSNPS and 107 sSNPs in the Mj sublineage, and 13 nsSNPs and 8 sSNPs in the Mn. These concatenated sequences were then compared with their respective ancestral sequences to obtain dN/dS. The raw counts of nonredundant nsSNPs and sSNPs pre- and postdiversification were compared for each sublineage using the *G* test, with *P* values shown. Analysis 1b: excluding singleton SNPs. The *G* test was based on 120 nsSNPs and 46 sSNPs for Mj and 8 nsSNPs and 4 sSNPs for Mn postdiversification. Analysis 2a: dN/dS was calculated for each isolate compared with its respective ancestral sequence (i.e., 153 Mj isolates were compared with the imputed ancestral sequence for Mj). Within each sublineage, the median dN/dS was calculated and is shown above. Analysis 2b: excluding singleton SNPs. The Wilcoxon signed-rank test was used to compare the median dN/dS postdiversification for each sublineage with its respective prediversification estimate.
*Fisher's exact test due to cell counts <5.

intact *pks15/1* locus of strain HN878 (Beijing genotype) induces hyperlethality in murine disease models (26), potentially explaining the emergence of the Beijing strain in a number of settings worldwide (27). Furthermore, compared with other clinical strains, strains 1471 and HN878 (Beijing genotypes) result in increased macrophage necrosis (28) and more progressive pathology in experimental infections (29). However, although certain strains have a propensity to cause accelerated life-threatening pathology in experimental models, it is not yet clear whether this property predicts epidemiologic success, as a strain that causes chronic, nonprogressive pathology may be the most likely to transmit.

In Nunavik, we observed a set of related strains that meet the epidemiologic criterion of success, without any clear genomic indicators of increased bacterial virulence. Instead, for the Mj sublineage, we observe an enrichment of nsSNPs since its introduction into this region, some of which are expected to affect the function of proteins that contribute to the survival of *M. tuberculosis* during infection. There are a number of potential causes of an increased dN/dS, including insufficient time for purifying selection to act, positive selection, relaxed purifying selection, and genetic drift. Whereas an increased dN/dS at the tips of a phylogenetic tree may indicate insufficient time for purifying selection (13), the postdiversification inflation of dN/dS holds even with the exclusion of evolutionarily recent singleton SNPs. Therefore, a simple time dependence is unlikely to be the only explanation. Positive selection is unlikely to inflate the dN/dS across the entire genome but rather should target genes with specific functions (15, 16). Although we did not identify any particular functional category of genes enriched in nsSNPs, this does not exclude positive selection on a small number of genes. However, it suggests that positive selection was not the pervasive force leading to a high dN/dS genome-wide. The remaining potential explanations for the dN/dS elevation are a genome-wide relaxation of purifying selection and genetic drift. The nsSNPs and deletions in putatively essential genes provide further support for these two interpretations.

The global *M. tuberculosis* population has been previously shown to evolve through mostly weak selection and strong drift (30); here we show that the same is true on a local level, to an even greater extent. Given that drift will have stronger effects when effective populations are reduced (31) and that our data suggest that population size remained more or less constant, we hypothesize that relaxation of purifying selection has contributed significantly to the evolution of the Nunavik strain of *M. tuberculosis*. Further investigation in this and other similar populations is needed. Regardless of the forces that have driven the elevated dN/dS, our findings suggest that *M. tuberculosis* has not thrived in Nunavik due to a unique virulence profile of the bacteria. It follows that *M. tuberculosis* control in this region, and in similar settings, will require looking beyond the bacterial culprit to the social conditions that foster TB.

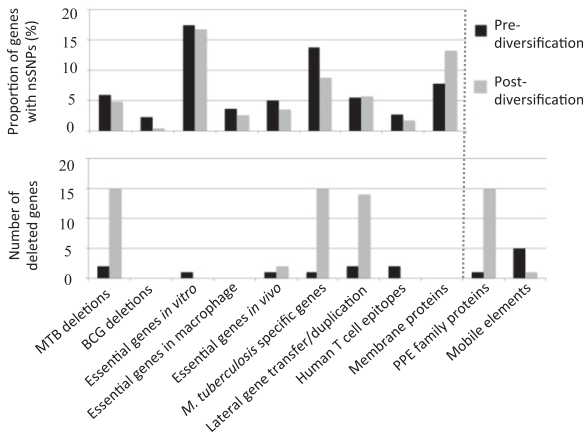


Fig. 4. Proportion of genes with nonsynonymous single-nucleotide polymorphisms (*Top*) and the number of deleted genes (*Bottom*) for the major sublineage, pre- and postdiversification. Gene categories are as defined in the publications: *M. tuberculosis* (MTB) deletions (36), bacillus Calmette–Guérin (BCG) deletions (37), essential genes in vitro (20), in macrophages (38), or in vivo (19), *M. tuberculosis*-specific genes (39), lateral gene transfer or duplication acquisition (39), human T-cell epitopes (7), genes coding membrane proteins (40), mobile elements (7), and genes coding PPE family proteins (7). Genes designated as PE/PGRS, PPE, or mobile elements were excluded from the SNP analysis (7).

Materials and Methods

Detailed methods can be found in *SI Materials and Methods*. In brief, the Nunavik region is composed of 14 Inuit communities, with a total population of 12,090 (in 2011). Between 1990 and 2013, there were 200 cases of TB in Nunavik, of which 163 were available for whole-genome sequencing using the MiSeq 250 System (Illumina). Reads were assembled and compared as previously described (32). The final dataset of SNPs excluded those in PE/PGRS and PPE genes, as well as mobile elements, as these may be prone to false positives (5, 7). Deletion events were identified with the Integrative Genomics Viewer (33) and confirmed by PCR and Sanger sequencing. Concatenated sequences of the SNPs were used to generate phylogenetic trees via the maximum likelihood method in Molecular Evolutionary Genetics Analysis [MEGA (34)]. Divergence times for the 163 Nunavik isolates were estimated using Bayesian Markov chain Monte Carlo methods [Bayesian Evolutionary Analysis by Sampling Trees (10, 11)], with H37Rv used as an outgroup.

We used three approaches to derive MRCAs. Using the concatenated sequences of SNPs across the 163 genomes, we first conducted an analysis that incorporated prior knowledge of the substitution rate of *M. tuberculosis* in the form of a calibration node for the Mj sublineage (analysis 1). We then performed an analysis agnostic to the reported substitution rate (i.e., without calibration), also using concatenated sequences (analysis 2). We then repeated this second analysis but applied a correction for the constant sites across the genomes (analysis 3).

Different coalescent models were tested to explore changes in effective population size over time (35). The AICM (14) was used to select the model providing the best fit. Bayesian skyline plots were generated (Fig. S2).

To calculate the dN/dS ratios, the ancestral sequences for each MRCA (Mj–Mn, Mj, and Mn) were reconstructed manually (Dataset S2). We then calculated the dN/dS pre- and postdiversification for the Mj and Mn sublineages, using both a phylogenetics-based approach (analysis 1) and a pairwise dN/dS analysis (analysis 2) (7). For both analyses, we repeated the dN/dS calculations after excluding SNPs that were present only once across all 163 genomes (singletons).

Ethical approval for this work was obtained from the McGill University Faculty of Medicine Institutional Review Board.

ACKNOWLEDGMENTS. The authors thank the Nunavik Regional Board of Health and Social Services for their collaboration on this study and Drs. Erwin Schurr, PhD and Michael Reed, PhD of the Research Institute of McGill University Health Centre for their input into the genetic analysis. This work was supported by the Canadian Institutes of Health Research (MOP 125858 to M.A.B. and D.M.) and Fonds de Recherche Santé Québec (29836 and 26274 to N.R.). B.J.S. was supported by the Canada Research Chairs Program (CRC 2289986).

1. Gagneux S, Small PM (2007) Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development. *Lancet Infect Dis* 7(5):328–337.
2. Wirth T, et al. (2008) Origin, spread and demography of the *Mycobacterium tuberculosis* complex. *PLoS Pathog* 4(9):e1000160.
3. Gagneux S, et al. (2006) Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *Proc Natl Acad Sci USA* 103(8):2869–2873.
4. Nguyen D, et al. (2003) Tuberculosis in the Inuit community of Quebec, Canada. *Am J Respir Crit Care Med* 168(11):1353–1357.
5. Roetzer A, et al. (2013) Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: A longitudinal molecular epidemiological study. *PLoS Med* 10(2):e1001387.
6. Marmiesse M, et al. (2004) 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(Pt 2):483–496.
7. Comas I, et al. (2010) Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. *Nat Genet* 42(6):498–503.
8. Coll F, et al. (2014) A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. *Nat Commun* 5:4812.
9. Walker TM, et al. (2013) Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: A retrospective observational study. *Lancet Infect Dis* 13(2):137–146.
10. Bouckaert R, et al. (2014) BEAST 2: A software platform for Bayesian evolutionary analysis. *PLOS Comput Biol* 10(4):e1003537.
11. Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* 29(8):1969–1973.
12. McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* 351(6328):652–654.
13. Rocha EPC, et al. (2006) Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J Theor Biol* 239(2):226–235.
14. Baele G, et al. (2012) Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Mol Biol Evol* 29(9):2157–2167.
15. Novichkov PS, Wolf YI, Dubchak I, Koonin EV (2009) Trends in prokaryotic evolution revealed by comparison of closely related bacterial and archaeal genomes. *J Bacteriol* 191(1):65–73.
16. Shapiro BJ, Alm EJ (2008) Comparing patterns of natural selection across species using selective signatures. *PLoS Genet* 4(2):e23.
17. Supply P, et al. (2013) Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of *Mycobacterium tuberculosis*. *Nat Genet* 45(2):172–179.
18. Wang J, et al. (2015) Insights on the emergence of *Mycobacterium tuberculosis* from the analysis of *Mycobacterium kansasii*. *Genome Biol Evol* 7(3):856–870.
19. Sassetti CM, Rubin EJ (2003) Genetic requirements for mycobacterial survival during infection. *Proc Natl Acad Sci USA* 100(22):12989–12994.
20. Sassetti CM, Boyd DH, Rubin EJ (2003) Genes required for mycobacterial growth defined by high density mutagenesis. *Mol Microbiol* 48(1):77–84.
21. Raghavan M, et al. (2014) The genetic prehistory of the New World Arctic. *Science* 345(6200):1255832.
22. Higdon J (2010) Commercial and subsistence harvests of bowhead whales (*Balaena mysticetus*) in eastern Canada and west Greenland. *J Cetacean Res Manag* 11:185.
23. Bonesteel S (2006) *Canada's Relationship with the Inuit*, ed Anderson E (published under the authority of the Minister of Indian Affairs and Northern Development and Federal Interlocutor for Métis and Non-Status Indians, Ottawa, Canada).
24. Grygier PS (1994) *A Long Way from Home: The Tuberculosis Epidemic Among the Inuit* (McGill-Queen's Univ Press, Montreal).
25. Alonso H, et al. (2011) Deciphering the role of IS6110 in a highly transmissible *Mycobacterium tuberculosis* Beijing strain, GC1237. *Tuberculosis (Edinb)* 91(2):117–126.
26. Reed MB, et al. (2004) A glycolipid of hypervirulent tuberculosis strains that inhibits the innate immune response. *Nature* 431(7004):84–87.
27. Parwati I, van Crevel R, van Soolingen D (2010) Possible underlying mechanisms for successful emergence of the *Mycobacterium tuberculosis* Beijing genotype strains. *Lancet Infect Dis* 10(2):103–111.
28. Amaral EP, et al. (2014) Pulmonary infection with hypervirulent Mycobacteria reveals a crucial role for the P2X7 receptor in aggressive forms of tuberculosis. *PLoS Pathog* 10(7):e1004188.
29. Ordway D, et al. (2007) The hypervirulent *Mycobacterium tuberculosis* strain HN878 induces a potent TH1 response followed by rapid down-regulation. *J Immunol* 179(1):522–531.
30. Hershberg R, et al. (2008) High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. *PLoS Biol* 6(12):e311.
31. Kuo CH, Moran NA, Ochman H (2009) The consequences of genetic drift for bacterial genome complexity. *Genome Res* 19(8):1450–1454.
32. Lee RS, et al. (2015) Reemergence and amplification of tuberculosis in the Canadian Arctic. *J Infect Dis* 211(12):1905–1914.
33. Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief Bioinform* 14(2):178–192.
34. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol* 30(12):2725–2729.
35. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biol* 4(5):e88.
36. Tsolaki AG, et al. (2004) Functional and evolutionary genomics of *Mycobacterium tuberculosis*: Insights from genomic deletions in 100 strains. *Proc Natl Acad Sci USA* 101(14):4865–4870.
37. Mostowy S, Tsolaki AG, Small PM, Behr MA (2003) The in vitro evolution of BCG vaccines. *Vaccine* 21(27–30):4270–4274.
38. Rengarajan J, Bloom BR, Rubin EJ (2005) Genome-wide requirements for *Mycobacterium tuberculosis* adaptation and survival in macrophages. *Proc Natl Acad Sci USA* 102(23):8327–8332.
39. Stinear TP, et al. (2008) Insights from the complete genome sequence of *Mycobacterium marinum* on the evolution of *Mycobacterium tuberculosis*. *Genome Res* 18(5):729–741.
40. Osório NS, et al. (2013) Evidence for diversifying selection in a set of *Mycobacterium tuberculosis* genes in response to antibiotic- and nonantibiotic-related pressure. *Mol Biol Evol* 30(6):1326–1336.
41. Cingolani P, et al. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w^{1118}; iso-2; iso-3. *Fly* 6(2):80–92.
42. Rutherford K, et al. (2000) Artemis: Sequence visualization and annotation. *Bioinformatics* 16(10):944–945.
43. Waddell PJ, Steel MA (1997) General time-reversible distances with unequal rates across sites: Mixing gamma and inverse Gaussian distributions with invariant sites. *Mol Phylogenet Evol* 8(3):398–414.
44. Tamura K (1992) Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Mol Biol Evol* 9(4):678–687.
45. Saitou N, Nei M (1987) The Neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4(4):406–425.
46. Felsenstein J (1985) Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39(4):783–791.
47. Comas I, et al. (2013) Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. *Nat Genet* 45(10):1176–1182.
48. Steenken W, Oatway WH, Petroff SA (1934) Biological studies of the tubercle bacillus: III. Dissociation and pathogenicity of the R and S variants of the human tubercle bacillus (H(37)). *J Exp Med* 60(4):515–540.
49. Rambaut A, Suchard M, Xie D, Drummond AJ (2014) Tracer v1.6. Available at beast.bio.ed.ac.uk/Tracer.
50. Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* 3(5):418–426.

APPENDIX 3-2

Supplementary Data:

Lee RS, Radomski N, Proulx J-F, Levade I, Shapiro BJ, McIntosh F, Soualhine H, Menzies D, Behr MA. Population genomics of *Mycobacterium tuberculosis* in the Inuit. *Proc Natl Acad Sci USA* 2015;112(44):13609-13614

Population genomics of *Mycobacterium tuberculosis* in the Inuit

Robyn S. Lee [a,b,c,1], Nicolas Radomski [b,c,1], Jean-Francois Proulx [d], Ines Levade [e], B. Jesse Shapiro [e], Fiona McIntosh [b,c], Hafid Soualhine [f], Dick Menzies [b,c,g], Marcel A. Behr [b,c,2]

Supporting Information

MATERIALS AND METHODS

Study population

The Nunavik region is 443,685 km^2 in size and is comprised of 14 Inuit communities, with a total population of 12,090 (Statistics Canada, 2011). Each of these communities is separated from the nearest village by a median distance of 137 km (IQR 110-178), without adjoining roads.

Bacteria

All specimens from TB suspects in Nunavik are sent to the mycobacteriology laboratory of the McGill University Health Centre (MUHC) for processing. Culture-positive specimens are then forwarded to the Laboratoire de Santé Publique du Québec for drug susceptibility testing. Between 2001 and 2013, there were 149 cases of microbiologically confirmed TB in Nunavik. All available isolates were included in this study and were provided by these two laboratories. Between 1990 and 2000, there were 51 cases of TB in Nunavik. 26 isolates were available from a previous study for these years (4).

DNA extraction

M. tuberculosis DNA was isolated as previously described in Lee *et al.* 2015 (32).

Whole Genome Sequencing

High throughput sequencing of extracted DNA was performed by the McGill University and Génome Québec Innovation Centre. The amount of gDNA was checked by Quant-iT™ PicoGreen® dsDNA Assay Kit (Life Technologies). gDNA was then fragmented by sonication using a TruSeq gDNA Library automate (Illumina). gDNA quality was estimated by High Throughput Quality Check for Massively Parallel Sequencing Library and fragments were multiplexed by 24 for paired-ends 250 base pair (bp) sequencing using MiSeq 250 System (Illumina).

Identification, annotation and confirmation of SNPs and deletions

Reads were aligned to H37Rv (NCBI Accession number NC_000962.3) and compared as previously described (32). Aligned reads were deposited in the National Center for

Biotechnology Information's Sequence Read Archive, under Accession number SRP039605 (BioProject PRJNA240330). Due to their repetitive nature, it is more difficult to accurately map reads to the PE/PGRS and PPE genes, as well as mobile elements, therefore these regions may be at higher risk of false positives (5, 7). Consequently, SNPs in these regions were excluded from the analyses presented in this manuscript. For a list of non-synonymous SNPs that were excluded, see Dataset S3. Additional analyses including SNPs in these regions did not alter our key findings.

Deletion events were identified against the H37Rv and CDC1551 reference genomes with Integrative Genomics Viewer (version 2.3.34) (33) and confirmed by PCR and Sanger sequencing. The SNPs and deletions were annotated against two reference genomes (uid57777 and uid57775 databases for H37Rv and CDC1551, respectively) using snpEff (version 3.3h) (41) and Artemis (version 15.0.0) (42), respectively.

Phylogenetic analysis

Concatenated sequences of the SNPs were used to generate phylogenetic trees via the Maximum Likelihood method in Molecular Evolutionary Genomics Analysis (MEGA, version 6) (34). Based on the lowest Bayesian Information Criterion (BIC), the General Time Reversible model (43) with uniform site-specific rate variation and Tamura 3-parameter (44) model were used to construct phylogenetic trees of genomes from Nuvavik and isolates from lineages 1, 2, 3, 4, 5, 6 and 7. The initial trees for the heuristic search were performed by the Neighbor-Joining method and Maximum Composite Likelihood approach. The branches of the trees presenting the highest log likelihood were condensed at the 80% bootstrap value (45). The *M. tuberculosis* genomes of lineages 1 (SAMEA1877171, SAMEA1877209, SAMEA1877280), 2 (SAMEA1877068, SAMEA1877086, SAMEA1877286), 3 (SAMEA1877096, SAMEA1877124, SAMEA1877277), 4 (SAMEA1877192, SAMEA1877238, SAMEA1877276), 5 (SAMEA1877073, SAMEA1877165, SAMEA1877206), 6 (SAMEA1877101, SAMEA1877190, SAMEA1877233) and 7 (SAMEA1877077, SAMEA1877197, SAMEA1877216) were obtained from the European Nucleotide Archive ERP001731 (46).

Molecular dating

Divergence times for the 163 Nunavik isolates were estimated using Bayesian Markov Chain Monte Carlo methods (Bayesian Evolutionary Analysis Sampling Trees (BEAST), versions 1.8 and 2.1.3) (10, 11). As the Tamura 3-parameter model of nucleotide substitution was not available in BEAST, we utilized the General Time Reversible (GTR) model (43) which had the next lowest BIC ($\Delta = 9.03$).

Divergence dates were estimated using all 163 genomes dated with the year of isolation. H37Rv was used as an out-group (with a date of isolation of 1905) (47). To assess the robustness of our findings, we compared three different approaches to deriving the MRCAs. Using the concatenated sequences of SNPs across the 163 genomes, we first conducted an analysis that incorporated prior knowledge of the substitution rate of *M. tuberculosis* in the form of a calibration node for the Mj sub-lineage (analysis 1). We identified the SNP difference between the two most divergent isolates from the Mj sub-lineage as 72 SNPs. Assuming equal divergence, we then applied the previously reported substitution rate of 0.5 SNPs/genome/year to obtain a mean node age, while the standard deviation was calculated using the extremes of frequently reported confidence intervals (0.3 SNPs/genome/year and 0.7 SNPs/genome/year, respectively (5, 9). The resultant mean node age and its standard deviation were then used as a prior for the MRCA of the Mj sub-lineage. We then compared these results to an analysis agnostic to the reported substitution rate (i.e. without calibration), also using concatenated sequences (analysis 2). We then repeated this second analysis but applied a correction for the constant sites across the genomes (analysis 3).

The null hypothesis of one molecular clock across all branches was tested prior to analysis using the likelihood ratio test in MEGA and was rejected with $p < 0.05$. Therefore, all models used an uncorrelated relaxed lognormal clock set at 1.3×10^{-7} substitutions per site in the genome per year. For analyses 1 and 2, a coalescent constant population tree prior was used. Analysis 3 utilized a coalescent Bayesian skyline tree prior, as described below. All models were run using a Markov Chain Monte Carlo (MCMC) chain length of 200,000,000, with 10% burn in and sampling every 10,000 generations. Convergence was assessed in Tracer (version 1.6) (48) with evidence of adequate mixing and all parameters had an effective sample size > 190. Maximum clade credibility trees were generated using TreeAnnotator, with 10% burn in. Summaries of the

posterior densities were generated for nodes with a probability of at least 0.8. The 95% highest posterior densities were used to reflect uncertainty in our estimates.

Coalescent-based analyses

Different coalescent models were tested, including the constant population size, exponential growth (assuming a constant growth rate through time) and the Bayesian skyline plot demographic model (a general, nonparametric prior that enforces no particular demographic history (35)). The posterior simulation-based analogue of Akaike's information criterion (AICM) (13) implemented in Tracer 1.6 has been used to select the model providing the best fit to our data. The estimations of the AICM from 1,000 bootstrap replicates support that, among the different models, the Bayesian skyline model provides the better fit overall (marginally lower value of AIC), followed by the constant population size model (Table S2). Substitution rates were estimated for each sub-lineage in BEAST using a GTR model of nucleotide substitution and a coalescent Bayesian Skyline tree prior. To formally assess changes in population demographics over time, Bayesian skyline plots were generated in Tracer for the 163 genomes overall, as well as the Mj and Mn sub-lineages individually (Fig. S2).

Calculation of dN/dS ratios

The ancestral sequences for each MRCA (Mj-Mn; Mj; and Mn) were reconstructed manually (Dataset S2). We calculated the dN/dS between Mj-Mn and Mj by comparing these reconstructed codon sequences using the Nei-Gojobori (Jukes-Cantor) method (49) in MEGA (version 6), as has been previously done for *M. tuberculosis* (7, 29, 31). Similarly, to obtain a dN/dS for Mj-Mn versus Mn, we compared these two reconstructed sequences.

For dN/dS calculations post-diversification, two methods were applied. First, we calculated the dN/dS for each sub-lineage compared to its ancestral sequence using a phylogenetic-based approach (analysis 1). In this method, we generated a concatenated sequence of codons for each sub-lineage based on all SNP loci. SNPs that were occurred in more than 1 isolate only contributed once to this sequence; no SNPs were double-counted. These concatenated sequences (one for each of the Mj and Mn sub-lineages) were then compared to their respective imputed ancestral sequences to obtain the dN/dS post-diversification. Second, we conducted a pairwise

dN/dS analysis (analysis 2) (7). In this analysis, we calculated dN/dS for each isolate of the Mj sub-lineage compared to the reconstructed ancestral sequence of Mj. We then determined the median dN/dS across the 153 pairwise comparisons. For both analyses 1 and 2, we repeated the dN/dS calculations after excluding SNPs that were present only once across all 163 genomes (singletons).

Comparisons to gene categories from the literature

nsSNPs were mapped to gene categories described in the literature: *M. tuberculosis*-specific genes (39), *in vitro* (19) or *in vivo* (18) essential genes, genes coding membrane proteins (40), lateral gene transfer or duplication acquisition (39), regions of differentiation (36), macrophage survival (38), human T cell epitopes (7) and deleted genes in BCG (37). In addition to these gene categories, deletions were also mapped to genes coding PE/PPE family proteins (7) and mobile elements (7). These comparisons were performed using the most recent version of the H37Rv reference genome (NCBI Accession number NC_000962.3) including gene nomenclatures of older versions of H37Rv used by these authors at the time of their studies.

Statistical analyses.

The distributions of pairwise SNPs within and between villages were compared using the Wilcoxon-Mann-Whitney test. The two-sample z test for difference in proportions was used to compare proportions of pairwise comparisons less than the pre-specified threshold, as a proxy of transmission during the study period. This test was also used to compare the proportions of genes with nsSNPs in each gene category pre- and post-diversification. For the dN/dS analyses, the raw numbers of nsSNPs and sSNPs pre and post-diversification were compared using the G-test (12) while the median pairwise dN/dS values post-diversification for each sub-lineage were compared to their respective pre-diversification value using the Wilcoxon signed rank test. All tests were two-tailed, with a p value of < 0.05 considered statistically significant. Analyses were conducted in Stata (v.13, StataCorp 2013) and R (v.3.1.2, available at https://cran.r-project.org/).

Ethics

Ethical approval for this work was obtained from the McGill University Faculty of Medicine Institutional Review Board. Individual patient consent was not required.

LEGENDS

Dataset S1: Identification by whole genome sequencing (on the left) and confirmation by PCR and Sanger sequencing (on the right) of deletions identified across 163 genomes from Nunavik, according to H37Rv (NCBI Reference Sequence: gi|448814763|ref|NC_000962\.3) and CDC1551 (NCBI Reference Sequence: gi|50953765|ref|NC_002755\.2) reference genomes.

Dataset S2: Phylogenetic clusters defined by patterns of single nucleotide polymorphisms (SNPs) identified in 163 genomes of *M. tuberculosis* from Nunavik.

Dataset S3: Comparison between mutations (single nucleotide polymorphisms and deletions) detected in Major (Mj) pre- and post-introduction diversification in Nunavik, according to gene categories: *M. tuberculosis* deletions (36), BCG deletions (37), essential genes *in vitro* (19), in macrophages (38) or *in vivo* (18). *M. tuberculosis*-specific genes (39), lateral gene transfer or duplication acquisition (39), human T cell epitopes (7), genes coding membrane proteins (40), mobile elements (7) and genes coding PPE family proteins (7). SNPs in mobile elements and genes encoding PE and PPE proteins were excluded from analyses (7).

Dataset S4: Non-synonymous single nucleotide polymorphisms (nsSNPs) of the major (Mj) sub-lineage pre- and post-diversification identified in 153 genomes of *M. tuberculosis* from Nunavik.

Dataset S5: Non-synonymous single nucleotide polymorphisms (nsSNPs) identified in essential genes of the major (Mj) sub-lineage post-diversification identified in 153 genomes of *M. tuberculosis*.

Figure S1: Most recent common ancestors (MRCAs) and deletion events during the evolution of major (Mj) sub-lineage of Nunavik. As similar estimates with overlapping 95% highest posterior density intervals were obtained in all 3 MRCA analysis, MRCA dates obtained via analysis 1 are presented for simplicity. Arrows indicate the positions and annotations of the H37Rv reference genome. A phylogenetic cluster was defined as at least 2 genomes sharing a minimum of two single nucleotide polymorphic loci. Grey and colored circles represent the common ancestors and phylogenetic clusters based on identified SNPs, respectively. Isolate (58385) had a unique single nucleotide polymorphism profile (i.e. was not clustered) and, as the posterior density for the divergence date of this isolate was <0.8, has not been shown.

Figure S2: Bayesian Skyline plots of *M. tuberculosis* in Nunavik. (A) All sub-lineages together; (B) The Major sub-lineage; (C) the Minor sub-lineage. The estimated effective population sizes through time are shown (black line). The shaded area represents the 95% credibility intervals.

Table S1: Model comparison by posterior simulation-based analogue of Akaike's information criterion (AICM, (13)). Lower AICM values indicate better model fit. The differences between AICM are reported. Positive values indicate better relative model fit of the row’s model compared to the column’s model.

Figure S1

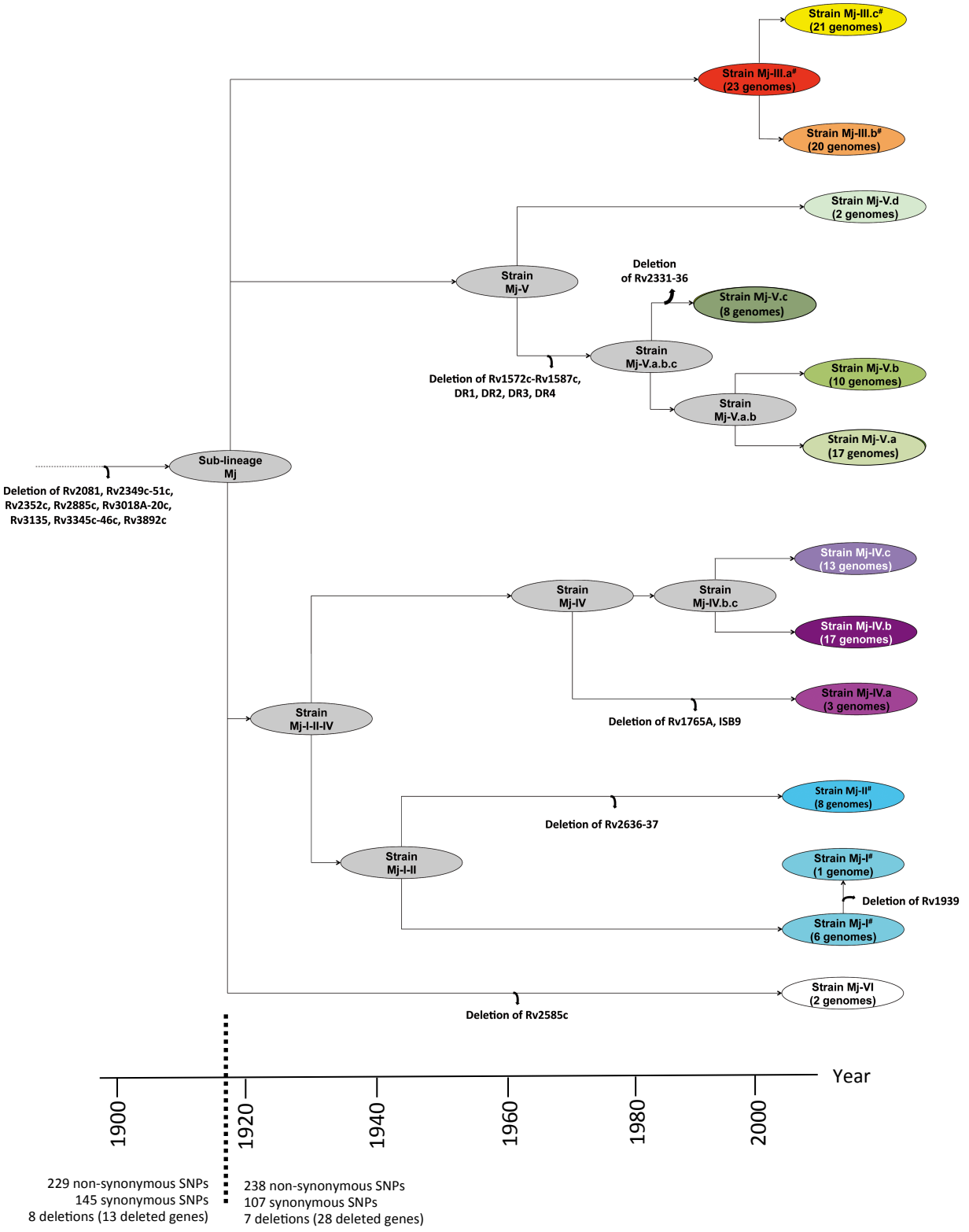
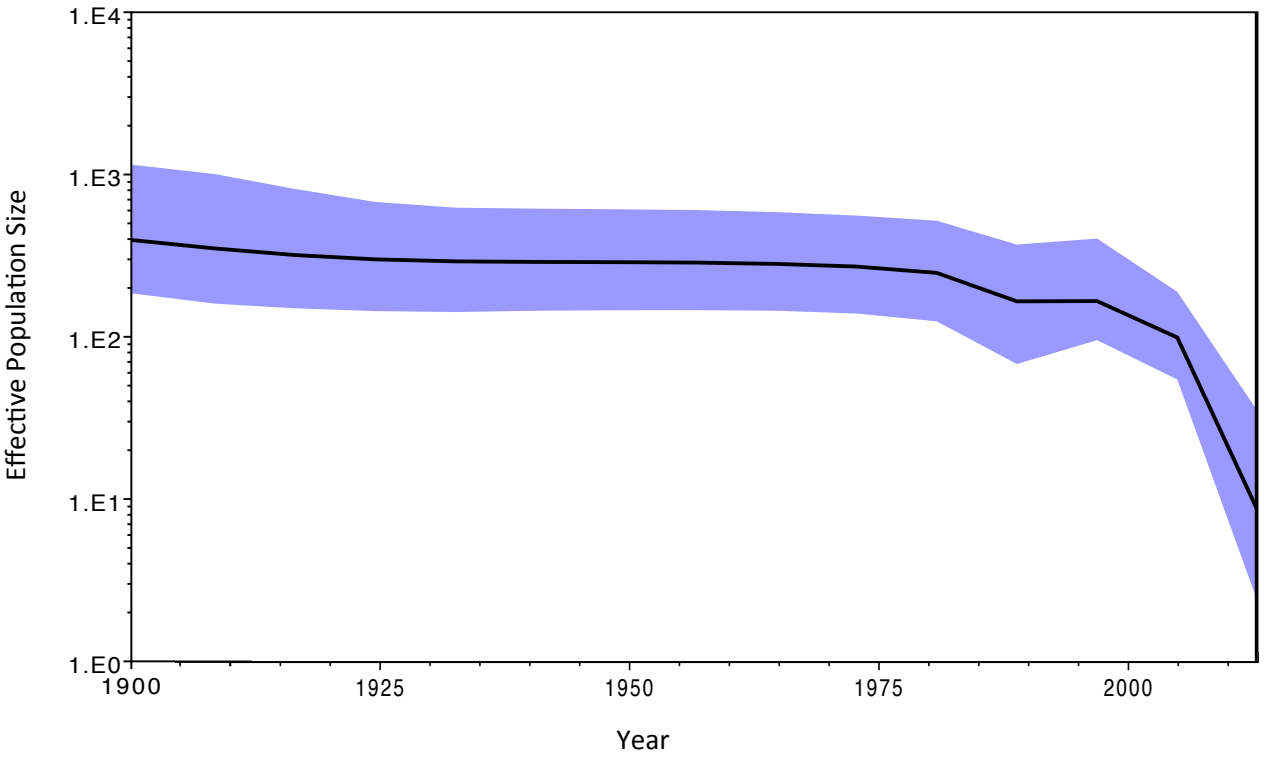
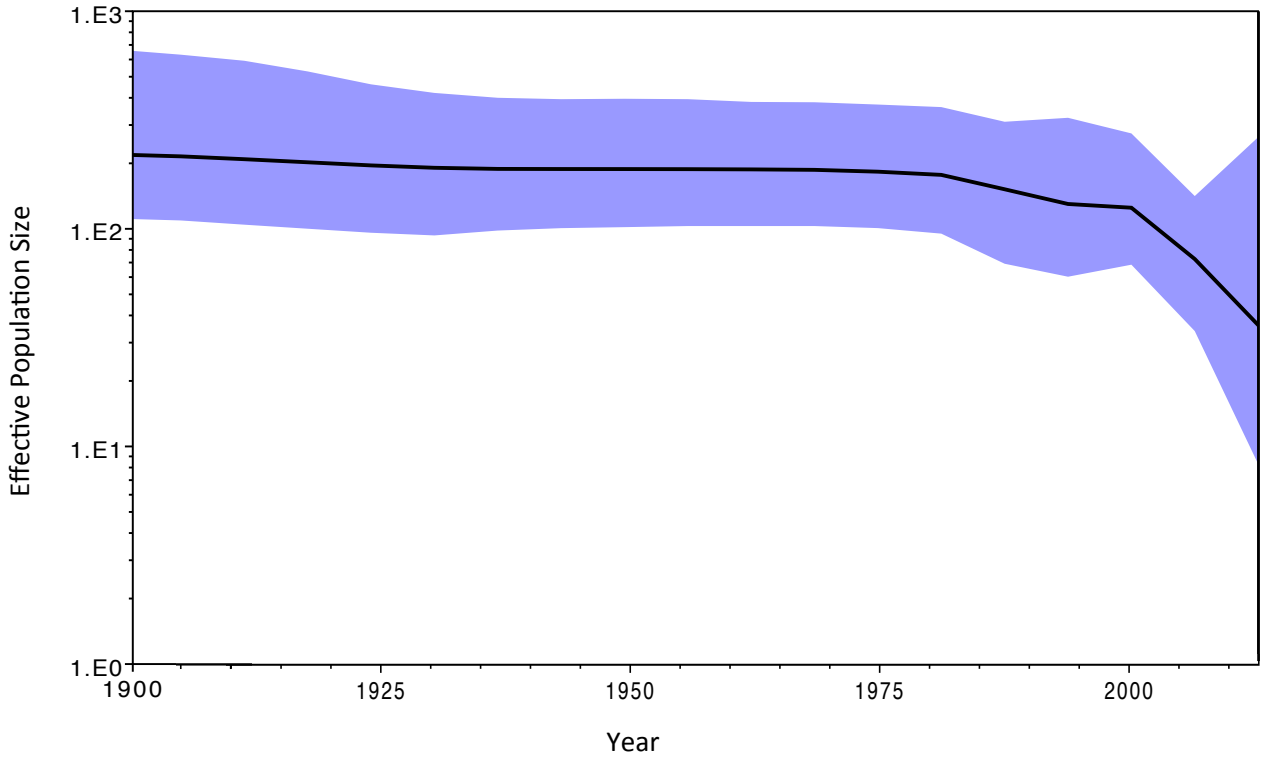

Strain Mj-III.c# (21 genomes)
Strain Mj-III.a# (23 genomes)
Strain Mj-III.b# (20 genomes)
Strain Mj-V.d (2 genomes)
Strain Mj-V
Deletion of Rv2331-36
Strain Mj-V.c (8 genomes)
Strain Mj-V.a.b.c
Deletion of Rv1572c-Rv1587c, DR1, DR2, DR3, DR4
Strain Mj-V.b (10 genomes)
Strain Mj-V.a.b
Strain Mj-V.a (17 genomes)
Sub-lineage Mj
Deletion of Rv2081, Rv2349c-51c, Rv2352c, Rv2885c, Rv3018A-20c, Rv3135, Rv3345c-46c, Rv3892c
Strain Mj-IV.c (13 genomes)
Strain Mj-IV
Strain Mj-IV.b.c
Strain Mj-IV.b (17 genomes)
Strain Mj-IV.a (3 genomes)
Deletion of Rv1765A, ISB9
Strain Mj-I-II-IV
Strain Mj-II# (8 genomes)
Deletion of Rv2636-37
Strain Mj-I-II
Strain Mj-I# (1 genome)
Deletion of Rv1939
Strain Mj-I# (6 genomes)
Strain Mj-VI (2 genomes)
Deletion of Rv2585c
Year
1900
1920
1940
1960
1980
2000
229 non-synonymous SNPs
145 synonymous SNPs
8 deletions (13 deleted genes)
238 non-synonymous SNPs
107 synonymous SNPs
7 deletions (28 deleted genes)

Figure S2

A. All *M. tuberculosis* sub-lineages of Nunavik



B. The Major sub-lineage of Nunavik



C. The Minor sub-lineage of Nunavik

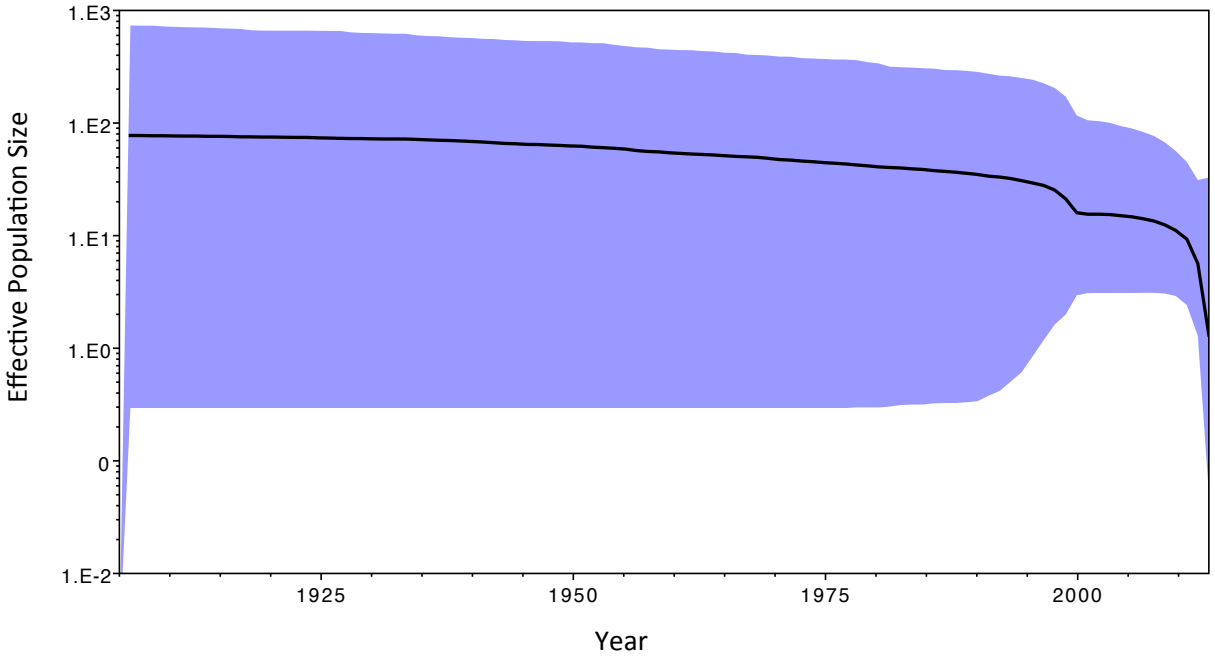


Table S1: Model comparison by posterior simulation-based analogue of Akaike's information criterion (AICM) (13).

Clock model	**Coalescent prior**	**AICM**	**SE**	**Exponential growth**	**Constant population size**	**Bayesian skyline**
Relaxed	Exponential growth	11830577.14	+/- 0.099	-	-10.577	-19.503
Relaxed	Constant population size	11830566.57	+/- 0.121	10.577	-	-8.926
Relaxed	Bayesian skyline	11830557.64	+/- 0.12	19.503	8.926	-

Lower AICM values indicate better model fit. The differences between AICM are reported. Positive values indicate better relative model fit of the row's model compared to the column's model.

APPENDIX 4

Supplementary data:

Lee RS, Proulx J-F, Menzies D, Behr MA. Progression to tuberculosis disease increases with multiple exposures. Under review at *Eur Respir J.*

Progression to Tuberculosis Disease Increases with Multiple Exposures

Robyn S. Lee, Jean-François Proulx, Dick Menzies,

Marcel A. Behr

Online supplementary material

Detailed Methods

Clinical assessment and management of contacts

As part of contact investigation, contacts were identified and evaluated for latent TB infection / active TB disease. Standardized data collection tools were used. Medical history was obtained by chart abstraction, supplemented by patient interview, and contacts were asked about TB symptoms. All contacts underwent a medical examination, a tuberculin skin test (TST) for those without prior TB infection and chest radiography if TST positive. Individuals with clinical suspicion of TB disease provided three spontaneous or induced sputum samples, which were sent for microscopy and mycobacterial culture. Individuals were diagnosed with active TB on the basis of growth of probe-confirmed *M. tuberculosis* ('confirmed TB'). Persons with clinical and radiographic findings consistent with TB, but absent culture confirmation, were classified as 'probable' TB.

Contacts identified with prevalent TB disease were treated accordingly, while those infected without disease were offered nine months isoniazid (INH) prophylaxis [1].

Newly-diagnosed ('new') TB infection

Individuals without a previous positive TST were tested as per the Canadian TB standards [1]. 5 units of purified protein derivative were injected intra-dermally, with induration measured 48-72 hours after injection using the ballpoint pen technique. If the initial TST was negative, the test was repeated at 8 weeks post-contact. In accordance

with contact investigation guidelines, a TST was considered positive if induration exceeded 5 mm [1].

A person was considered to have 'new' TB infection if he/she had a positive TST, either with no previous TST ('new positive' TST) or with a previously negative TST ('TST conversion').

As it is possible that an individual with a new positive TST was infected years in the past, we performed sensitivity analyses restricted to documented TST converters. To address the timing of those with a documented TST conversion, we examined the time of last negative test. Five of 29 (17%) cases had a previously negative test conducted within the year preceding the outbreak compared to 35 of 94 (37%) controls, arguing against the possibility that cases were more recently infected compared to controls.

Exposure ascertainment

Total exposures

We were able to tabulate total exposures by examining the contact lists provided by the 50 individuals with microbiologically-confirmed TB, including residence or shared attendance at community gathering houses. Each time a person was listed as a contact, this was counted as an exposure. These lists were obtained as part of the public health response, and were therefore only provided by those with active TB.

The precise intensity of contact (i.e., duration and frequency) was not included in modeling of exposure, as these data were not obtained in a consistent manner throughout the 'outbreak'. There were two main reasons for discrepancies in measurement of these variables. Firstly, given the high exposure intensity and short duration of this crisis, numerous individuals were repeatedly listed as contacts of persons with active TB. To avoid unnecessary burden on these individuals and the repetition of tuberculin skin tests / chest x-rays within a short period of time, they were not always re-assessed with each new contact. The decision to re-investigate with each new contact was made on a case-by-case basis by local health care providers. Secondly, a large number of contacts occurred at gathering houses. As these were identified by public health as potential sites of transmission, we have included contact at these houses between attendees as well as residents in our analyses, but the precise duration of exposure between individuals at such venues was not available.

Genotypic exposures

Genotypes were assigned to each of these individuals based on a previous molecular epidemiologic analysis of this crisis [2]. In brief, 49/50 of those with confirmed disease shared the same pattern on mycobacterial interspersed repetitive units (MIRU). Using whole genome sequencing and epidemiologic data, it was revealed that these were comprised of at least 6 different subgroups of transmission, with genotypes diverging from one another as early as 2007 (Figure 5 in [2]). These subgroups were closely-related with 6 cluster-defining single nucleotide polymorphisms separating the most genetically-divergent groups. These SNPs were bi-directional, with 3 in one subgroup and 3 in the

other, thus precluding transmission between these cases; SNPs do not revert to wild-type in *M. tuberculosis* [3].

One isolate was unavailable for sequencing. As this isolate had a non-identical MIRU pattern compared to the other 49 from 2011-2012, it was therefore was considered a unique genotype for this analysis – yielding a potential maximum of seven different genotypic exposures.

We assigned links based on epidemiologic data (i.e., identified during interviews). Links were not assigned between patients based on genotypic data alone, as suggested by previous studies (e.g., [4, 5]). Furthermore, such assignment would likely have introduced bias in favour of increased exposures for the cases, as genotypic data was not available for controls. As the molecular analysis was conducted retrospectively, it was also not feasible to re-interview individuals to assess for potentially missed contact based on genotyping results. Similarly, this would have been expected to a differential bias in favour of the cases.

Covariates

Covariate data was collected as part of routine contact investigation. These included address of residence, age at infection (based on first documented positive TST), sex, cigarette smoking, Bacillus Calmette-Guerin (BCG) vaccination and previous medical history including HIV and relevant comorbidities. The number of persons per room and

residing with a person with smear positive disease were determined retrospectively, with details below.

Current cigarette smoking

This was assessed at the time of contact investigation and was included as a binary variable. Data on intensity (e.g., number of cigarettes smoked per day) were missing for 41% of those investigated, exceeding that which can be meaningfully imputed.

HIV and other comorbidities

HIV testing was performed for all individuals with diagnosis of active TB. Among those with recent infection, one person diagnosed with active TB was previously known to be HIV positive. As no other relevant comorbidities (e.g., diabetes, renal dysfunction, cancer or other immunosuppressive disorders) were found among those with recent infection, these variables were not included in regression models.

The number of persons per room

This variable was used as a measure of occupancy and was calculated for each dwelling as the number of persons residing in a house divided by the number of rooms. These data were obtained/calculated as follows:

The number of persons residing at each residence, *regardless* of infection/disease status, was tabulated using a population-level database provided by public health. During the 'outbreak', clinical files from the village nursing station were used to obtain a complete

listing of current residents in the community. Addresses of residence were obtained by cross-referencing with a community-wide census provided by the village council for this purpose. If discrepancies were noted between these addresses and contact investigation data, addresses were updated using the latter. If an individual's address was missing from contact investigation data or the village census, efforts were made to obtain these data via consultation with patients directly if possible or alternatively, with local clinic support staff.

The number of bedrooms per dwelling were provided by the Katavik Municipal Housing Bureau for houses built up before 2012, and supplemented with housing assessment data from [6]. The number of rooms per dwelling were therefore calculated as the number of bedrooms plus one room. This additional room is the equivalent of the kitchen/dining room and living area, as these are open-concept in all dwellings in this community. In accordance with Statistics Canada, storage rooms, vestibules, bathrooms and hallways were not counted as additional rooms [7].

Residing with a person with smear positive disease

This was determined on a per subject basis, to assess the potential impact of household contact with highly contagious smear positive individuals. If an individual had smear positive disease, but did not reside with another person with smear positive disease, (s)he was assigned a value of 0 for this covariate, as one could not auto-contribute to risk of disease.

Additional potential TB determinants

Alcohol consumption (yes/no) was missing for 85/149 (57%) of those with recent infection and was therefore not included in our analyses. Data on nutritional status (e.g., body mass index or dietary habits) were not collected as part of routine contact investigation and were therefore not available for inclusion. However, a recruitment-based case-control study in this same community in 2013 found no association between these factors (such as measured serum micronutrient levels, reported nutritional intake, body mass index, and alcohol consumption) and progression from those with recent infection to disease [8].

Missing data

Three controls had missing addresses and were therefore excluded from analyses. Percentages of missing data for covariates were low (Table S2).

Missing data were estimated using multiple imputation with chained equations (m=500), under a Missing At Random assumption. Separate imputation datasets were generated for analysis 1 (contact with any potential source, new infection) and analysis 2 (contact with persons with smear positive disease only, new infection). Analyses 1b and 2b were conducted using subsets of these.

As the principles guiding development of imputation models and diagnostics are the same throughout, these have been described only for analysis 1a. At a minimum, imputation models for each variable included all covariates to be considered in the analysis model

(age at infection, sex, persons per room, current smoking, residing with a person with smear positive disease). An interaction term was considered in the analysis for persons per room and residing with a person with smear positive disease, but did not require imputation as both of the subjects with missing data on persons per room were known *not* to reside with a person with smear positive disease, making this variable equal to 0. Finally, the outcome variable was included in all imputation models.

The inclusion of auxiliary variables in addition to the above was based on whether these predicted the imputation variable and/or its missingness, as assessed using linear or logistic regression on the complete data and whether inclusion of this auxiliary variable improved model fit. For smoking, such auxiliary variables included BCG vaccination status and presence of cough at time of contact investigation. BCG data was missing on three individuals; one was born after vaccination was discontinued in Nunavik, and was therefore coded as 0. The remaining 2 were imputed, as described, with smoking included as an auxiliary variable. Cough was missing for 19 individuals and was imputed. Only 5 of those with missing cough were also missing data on smoking, and missingness was not associated with being diagnosed with active TB.

For persons per room, the total number of people residing in the house and number of rooms were assessed as auxiliary variables; total residing in the house was a strong predictor and was subsequently included in the imputation model. To avoid implausible values, a truncated regression was used, with imputed values restricted to the range of the observed data.

Final imputation diagnostics included assessment of convergence for all parameters (iterations increased accordingly) and comparison of the observed versus imputed data. Imputed values were also assessed for plausibility. The full analysis model was used to evaluate Monte Carlo error, with repeatability considered acceptable as per the recommendations of [9].

Assessing linearity of all continuous variables

A visual inspection was first conducted using lowess regression to assess possible deviations from linearity. Fractional polynomials (FP) were fit for each continuous variable, with sequential selection. This included a log-transformation of these variables. With the exception of genotypic exposures for analysis 1a which was best modeled with a cubic function, analyses failed to reject the FP m=1 p=1 model. Therefore, all other continuous variables employed a linear functional form.

Regression analyses

To account for clustering by household, we used generalized estimating equations with a logit link and robust standard errors. Univariate analyses were conducted to examine the association between potential risk factors and progression to disease. Based on previously reported results [6], we evaluated for an interaction between residing with a person with smear positive disease and the number of persons per room; in order to maintain hierarchy, both of these variables were included in preliminary multivariate models regardless of significance on univariate analysis. An interaction term was considered

significant if $p<0.1$. Other covariates were assessed in preliminary multivariate analyses if p was <0.2 on univariate analysis. Final multivariate models were selected using the Quasi-Information Criterion (QICu) [10]. The QIC takes into consideration both sample size and the number of parameters therein, to avoid over-parameterizing the models. This is of particular importance given the small sample size, as inclusion of unnecessary or extraneous variables would result in over-fitting and reduce precision of all effect estimates.

References

1. Public Health Agency of Canada and Canadian Lung Association. Canadian tuberculosis standards, 7th Edition. 2014.

2. Lee RS, Radomski N, Proulx J-F, Manry J, McIntosh F, Desjardins F, Soualhine H, Domenech P, Reed MB, Menzies D, Behr MA. Reemergence and amplification of tuberculosis in the Canadian arctic. J Infect Dis 2015; 211: 1905–1914.

3. Perez-Lago L, Lirola MM, Herranz M, Comas I, Bouza E, Garcia-de-Viedma D. Fast and low-cost decentralized surveillance of transmission of tuberculosis based on strain-specific PCRs tailored from whole genome sequencing data: a pilot study. CMI 2015; 21: 249–249.

4. Braden CR, Templeton GL, Cave MD, Valway S, Onorato IM, Castro KG, Moers D, Yang Z, Stead WW, Bates JH. Interpretation of restriction fragment length polymorphism analysis of *Mycobacterium tuberculosis* isolates from a state with a large rural population. J Infect Dis 1997; 175: 1446–1452.

5. Bryant JM, Schurch AC, van Deutekom H, Harris SR, de Beer JL, de Jager V, Kremer K, van Hijum SAFT, Siezen RJ, Borgdorff M, Bentley SD, Parkhill J, van Soolingen D. Inferring patient to patient transmission of *Mycobacterium tuberculosis* from whole genome sequencing data. BMC Infect Dis 2013; 13: 110–110.

6. Ahmed Khan F, Fox GJ, Lee RS, Riva M, Benedetti A, Proulx JF, Jung S, Hornby K, Behr MA, Menzies D. Housing characteristics as determinants of tuberculosis in an Inuit community: a case-control study. 19th Annual Conference of The International Union Against Tuberculosis and Lung Disease - North America Region. Vancouver; 2015.

7. Statistics Canada. Rooms of private dwelling. www.statcan.gc.ca/eng/concepts/definitions/dwelling01. Date last updated: April 27 2015. Date last accessed: April 20 2016.

8. Fox GJ, Lee RS, Lucas M, Ahmad Khan F, Proulx J-F, Hornby K, Jung S, Benedetti A, Behr MA, Menzies D. Inadequate diet is associated with acquiring Mycobacterium tuberculosis infection in an Inuit community: A Case-Control Study. Annals ATS 2015; 12(8): 1153-1162.

9. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Statist Med 2011; 30: 377–399.

10. Pan W. Akaike's information criterion in generalized estimating equations. Biometrics 2001; 57: 120–125.

Tables

TABLE S1
Summary characteristics of the 50 persons with microbiologically confirmed active TB, 34 of which had recent infection

Characteristic	**No. (%)**
Age, in years *	
0-4.9	4 (8)
5-9.9	1 (2)
10-14.9	2 (4)
15-24.9	23 (46)
25-34.9	10 (20)
35-44.9	4 (8)
45+	6 (12)
Male sex	28 (56)
Cavity on chest x-ray	11 (22)
Sputum smear positive	11 (22)

* Age at treatment initiation.

TABLE S2
Missing data imputed, confirmed cases only

Variable	Analysis 1a (n=149)*	Analysis 2a (n=99)
No. (%) missing current smoking	21 (14)	14 (14)
No. (%) missing persons per room	2 (1)	0 (0)
No. (%) missing BCG	2 (1)	2 (2)

*23 children under the age of 10 were not asked by health care providers about cigarette smoking status and are considered non-smokers.

TABLE S3

Comparison of villagers with new infection and confirmed versus probable TB disease

Variables of interest	Confirmed TB disease (n=34)	Probable TB disease (n=17)	p value
Age at infection, median (interquartile range, IQR), y	19.5 (15.3-28.1)	5.6 (2.9-16.4)	0.002†
No. (%) under 5 y age	4 (12)	8 (47)	0.012#
No. (%) male sex	18 (53)	6 (35)	0.234‡
No. (%) current smoking	23 (77)	4 (25)	0.001‡
No. (%) vaccinated with Bacillus Calmette-Guerin (BCG)	25 (76)	7 (41)	0.016‡
No. (%) residing with a person with smear positive disease	7 (21)	6 (35)	0.256‡
No. (%) comorbidities (HIV, diabetes, renal dysfunction, other immunosuppressive disorders)	1 (3)	0 (0)	0.667#
Total exposures, median (IQR)	15 (3-23)	3 (1-4)	0.006†
Genotypic exposures, median (IQR)	5 (2-6)	2 (1-3)	0.007†
Persons per room, median (IQR)	1·8 (1·3-2·7)	2 (1.8-3)	0.236†

† Mann-Whitney ranksum test. ‡ Chi-square test with 2 degrees of freedom. # Fisher's Exact test. Non-missing data are used for the denominator of proportions. A two-sided p value of <0.05 is considered statistically significant.

TABLE S4
Genotypic exposures to any potential source and progression to active TB

		Univariate		Multivariate	
	Odds ratio	**95% CI**	**p value**	**Odds ratio**	**95% CI**
Analysis 1a – Contact with any potential source, newly diagnosed infection					
Age at infection	1.00	0.97-1.03	0.806	Not in final model	
Male sex	0.93	0.44-1.94	0.844	Not in final model	
Current smoking	1.61	0.58-4.51	0.360	Not in final model	
BCG	0.64	0.24-1.74	0.386	Not in final model	
Residing with a person with smear positive disease	2.03	0.67-6.14	0.208	Not in final model	
Genotypic exposures, **cubic**	1.01	1.00-1.01	<0.0005	1.01	1.00-1.01
Persons per room*	1.12	0.98-1.28	0.086	1.16	1.01-1.34
Analysis 1b - Contact with any potential source, tuberculin skin test conversion only					
Age at infection	1.01	0.98-1.04	0.593	Not in final model	
Male sex	0.72	0.31-1.67	0.442	Not in final model	
Current smoking	2.19	0.61-7.90	0.231	Not in final model	
BCG	1.30	0.33-5.21	0.707	Not in final model	
Residing with a person with smear positive disease	3.17	0.85-11.79	0.085	0.18	0.01-3.50
Genotypic exposures, **linear**	1.12	1.06-1.18	<0.0005	1.79	1.35-2.37
Persons per room*	1.13	1.00-1.28	0.056		
Persons per room* *when **not** residing with a person with smear positive disease*				1.12	1.00-1.27
Persons per room* *when residing with an person with smear positive disease*				1.54†	1.10-2.16

*For comparability to [6]. Persons per room scaled such that odds ratio corresponds to a 1 person increase in a 5-person house. †p=0.018 for interaction between persons per room and residing with a person with smear positive disease; this OR represents the joint effect of adding 1 person to a 5-person house when residing with an individual with smear positive disease. Age at infection and persons per room are centered at the overall mean for analysis 1a, at 20.8 years and 1.7 persons per room, respectively.

TABLE S5

Exposure to potential sources with smear positive disease only and progression to active TB

		Univariate		Multivariate	
	Odds ratio	**95% CI**	**p value**	**Odds ratio**	**95% CI**
Analysis 2a – Potential sources with smear positive disease only, newly diagnosed infection					
Age at infection	1.01	0.97-1.05	0.603	Not in final model	
Male sex	1.23	0.54-2.81	0.623	Not in final model	
Current smoking	1.26	0.36-4.38	0.719	Not in final model	
BCG	0.93	0.23-3.71	0.922	Not in final model	
Residing with a person with smear positive disease	1.40	0.44-4.43	0.572	Not in final model	
Genotypic exposures, linear	1.52	1.16-2.01	0.003	1.69	1.27-2.26
Persons per room*	1.24	1.09-1.41	0.001	1.31	1.13-1.51
Analysis 2b – Potential sources with smear positive disease only, tuberculin skin test conversion only					
Age at infection	1.01	0.97-1.06	0.574	Not in final model	
Male sex	1.00	0.40-2.52	1.000	Not in final model	
Current smoking	1.42	0.36-5.64	0.621	Not in final model	
BCG	1.28	0.23-6.95	0.778	Not in final model	
Residing with a person with smear positive disease	2.10	0.54-8.15	0.284	Not in final model	
Genotypic exposures, linear	1.60	1.20-2.14	0.001	1.80	1.35-2.40
Persons per room*	1.29	1.12-1.48	0.001	1.36	1.16-1.60

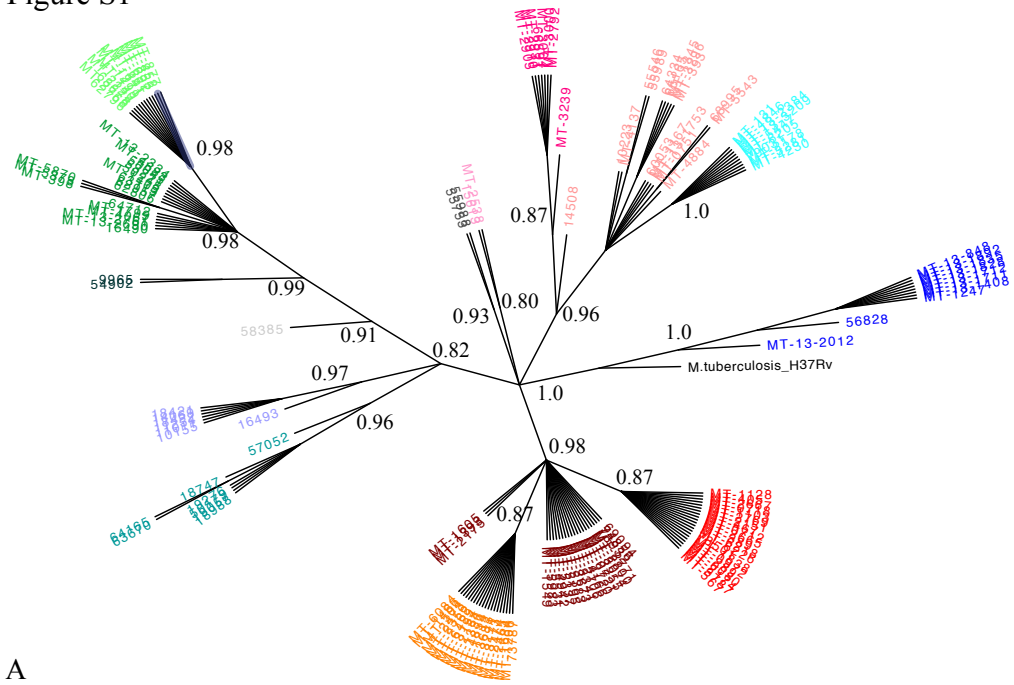
For comparability to [6]. *Persons per room scaled such that odds ratio corresponds to a 1 person increase in a 5 person house. Age and persons per room are centered at the overall mean for analysis 1a, at 20.8 years and 1.7 persons per room, respectively. Genotypic exposures centered at 1, as all individuals had at least 1 contact.

APPENDIX 5

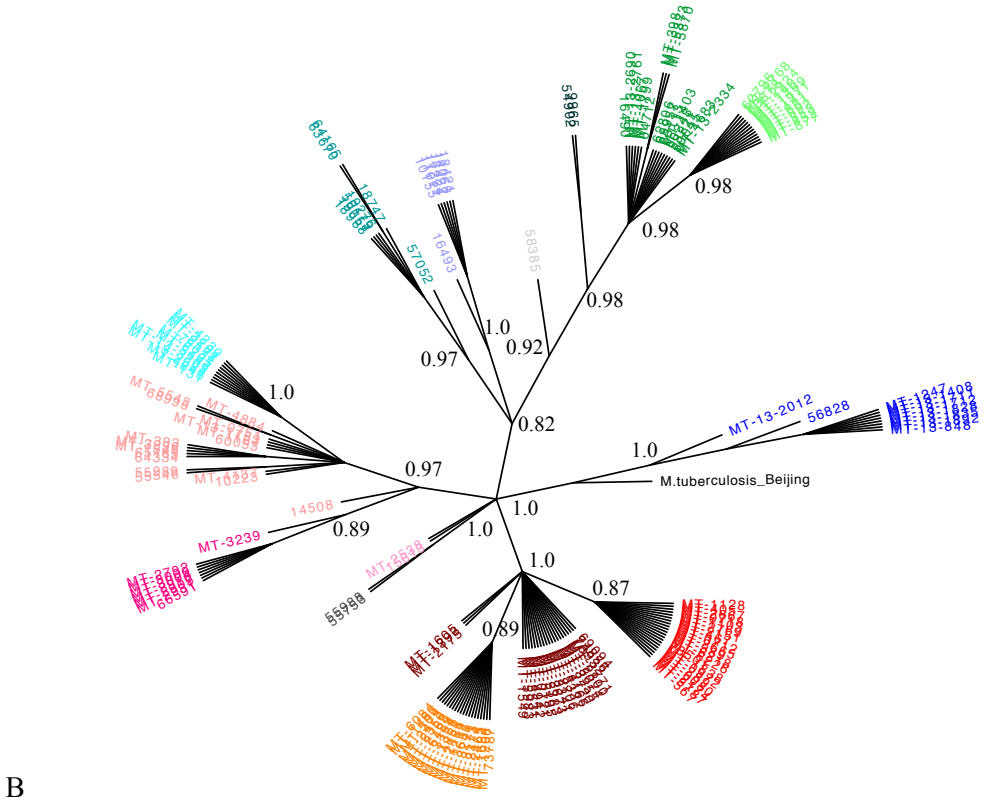
Supplementary data:

Lee RS and Behr MA. Does choice matter? Reference-based alignment for molecular epidemiology of tuberculosis. Accepted at *J Clin Micro* on April 5, 2016.

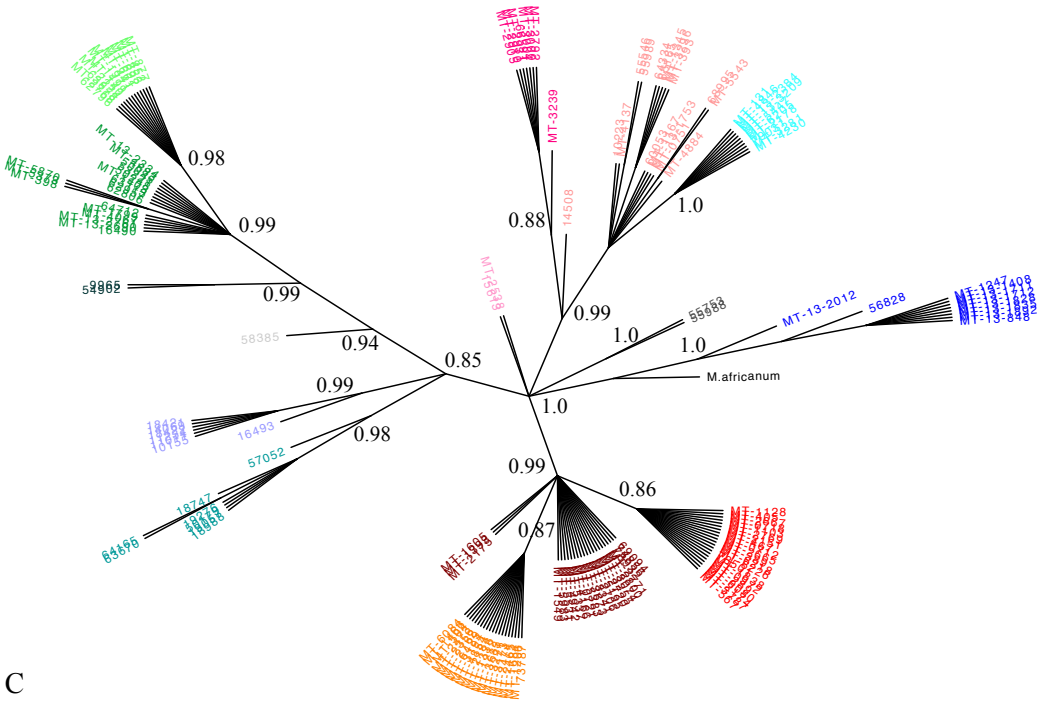
Figure S1


0.98
0.98
0.99
0.91
0.97
0.96
0.82
0.87
0.80
0.93
0.96
1.0
1.0
MT-3239
14508
MT-13-2012
56828
M.tuberculosis_H37Rv
16493
57052
58385
0.98
0.87
0.87

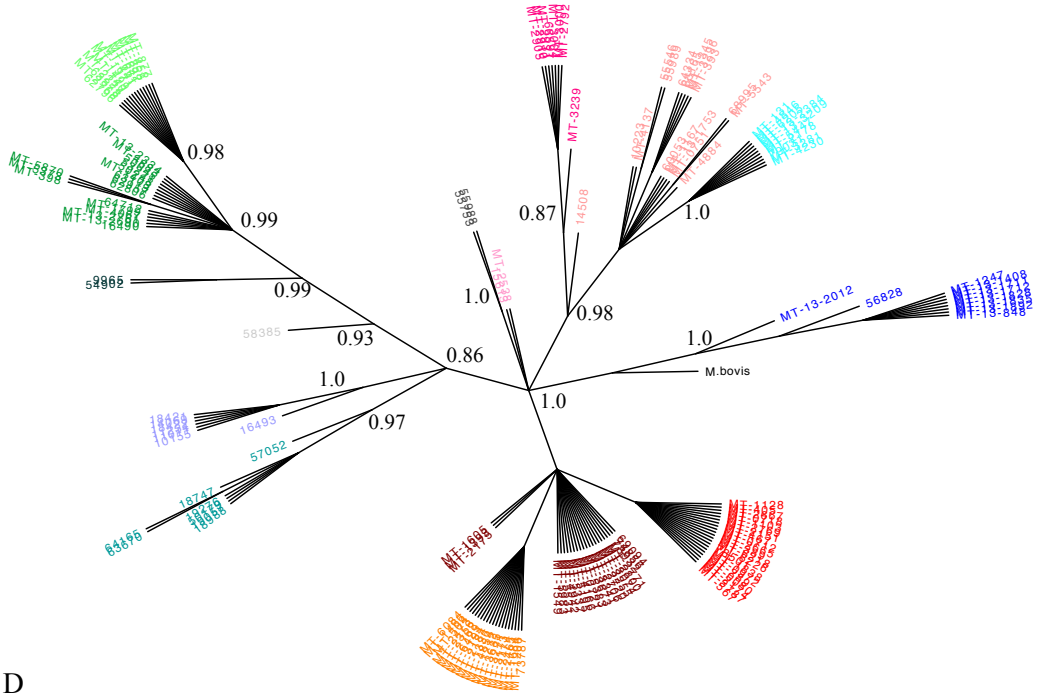
A


0.98
0.98
0.98
0.92
1.0
0.97
0.82
1.0
1.0
0.97
0.89
1.0
1.0
0.87
0.89
MT-13-2012
56828
M.tuberculosis_Beijing
MT-3239
14508
16493
57052
58385

B



C



D

Legend. Maximum likelihood trees with 1000 bootstrap replicates, with branches below an 80% bootstrap threshold collapsed (branch lengths are therefore not to scale). For clarity, bootstrap p values are indicated only for the most proximal node defining each cluster. Each tree was rooted on its respective reference. Isolates were coloured for their respective clusters identified according to CDC1551 (and H37Rv (1)). Isolates were then kept the same colour across all panels, to facilitate quick comparison between the new reference analysis and CDC1551. See Table S4 for cluster names. A – Reference *M. tuberculosis* Lineage 4 H37Rv, using the Tamura 3-parameter model (2) of nucleotide substitution with 1,405 SNP loci. B – Reference *M. tuberculosis* Lineage 2 CCDC5079, using the General Time Reversible (3) model of nucleotide substitution with 2,048 SNP loci. C – Reference *M. africanum* (Lineage 1), using the Tamura 3-parameter model of nucleotide substitution with 2,721 SNP loci. D – Reference *M. bovis*, using the GTR model of nucleotide substitution with 2,803 SNP loci.

TABLE S1 Average nucleotide identity (ANI) divergence between reference genomes [a]

		Reference genomes						
		Mycobacterium tuberculosis, lineage 4 H37Rv	*Mycobacterium tuberculosis,* lineage 4 CDC1551	*Mycobacterium tuberculosis,* lineage 2 CCDC5079	*Mycobacterium africanum* GN041182	*Mycobacterium bovis* AF2122/97	*Mycobacterium canettii* CIPT 140010059	*Mycobacterium kansasii* [b] ATCC 12478
Mycobacterium tuberculosis, lineage 4	H37Rv	-	0.10 (99.37)	0.12 (99.17)	0.18 (98.82)	0.19 (98.71)	0.85 (95.95)	19.01 (54.40)
Mycobacterium tuberculosis, lineage 4	CDC1551	0.11 (99.94)	-	0.16 (99.04)	0.18 (98.91)	0.21 (98.93)	0.85 (96.06)	19.02 (54.36)
Mycobacterium tuberculosis, lineage 2	CCDC5079	0.12 (99.68)	0.15 (99.83)	-	0.22 (98.71)	0.23 (98.35)	0.88 (96.06)	19.13 (54.36)
Mycobacterium africanum	GN041182	0.14 (99.27)	0.16 (99.62)	0.18 (99.34)	-	0.15 (99.02)	0.85 (95.98)	19.02 (54.39)
Mycobacterium bovis	AF2122/97	0.16 (99.18)	0.18 (99.44)	0.20 (99.00)	0.14 (99.61)	-	0.86 (95.69)	19.00 (54.36)
Mycobacterium canettii	CIPT 140010059	0.75 (96.94)	0.75 (97.07)	0.77 (96.95)	0.75 (97.09)	0.76 (96.67)	-	19.02 (54.39)
Mycobacterium kansasii	ATCC 12478	13.69 (16.25)	13.67 (16.10)	13.69 (16.09)	13.66 (16.13)	13.66 (16.19)	13.66 (16.26)	-

[a] Percent divergence is indicated, with the average percentage of nucleotides used in each analysis in brackets. Each genome was, in turn, fragmented into consecutive 500 base-pair segments and queried against all other complete reference genomes using jSpecies (v.1.2.1, (4)). Both BLASTn (5) and MUMmer (6) algorithms were applied. BLASTn settings were as in (7), requiring ≥70% identity over ≥70% of the alignment. Default settings were used for MUMmer. For each algorithm, the mean ANI was calculated (e.g., the mean of the pair A_{500} vs $B_{COMPLETE}$ and B_{500} vs $A_{COMPLETE}$ (8)). This was then used to tabulate ANI divergence (100% - the mean of pairwise ANI, (7)). ANI divergences calculated using BLASTn are indicated in red, while those using MUMmer are indicated in blue. In brackets, the average percent of total nucleotides in the query genome that were used in the calculation has been indicated for each comparison.
[b] pMK plasmid sequence not used for alignment.

TABLE S2 Genome coverage for each isolate

	Genome coverage (%) at 1x depth by reference genome						
Isolate	*M. tuberculosis*, lineage 4 H37Rv	*M. tuberculosis*, lineage 4 CDC1551	*M. tuberculosis*, lineage 2 CCDC5079	*M. africanum* GN041182	*M. bovis* AF2122/97	*M. canettii* CIPT 140010059	*M. kansasii* ATCC 12478
9965	98.86	99.25	98.75	98.88	99.27	95.04	34.63
10155	98.90	99.34	98.78	98.92	99.32	95.04	34.20
10223	98.70	99.13	98.80	98.94	99.19	95.12	34.90
11011	98.92	99.35	98.80	98.94	99.34	95.06	34.35
11234	98.83	99.25	98.71	98.84	99.26	94.99	33.66
14069	98.91	99.34	98.79	98.92	99.33	95.07	34.35
14508	98.71	99.12	98.79	98.92	99.19	95.11	34.84
15613	98.90	99.30	98.80	98.91	99.29	95.07	34.70
16490	98.82	99.25	98.70	98.84	99.25	95.01	34.07
16493	98.88	99.32	98.78	98.88	99.30	95.06	34.36
18421	98.84	99.26	98.73	98.86	99.28	95.03	33.83
18422	98.91	99.34	98.80	98.94	99.32	95.10	34.60
18747	98.86	99.28	98.77	98.90	99.29	95.03	34.21
18988	98.82	99.25	98.75	98.85	99.26	95.03	34.13
19057	98.90	99.31	98.80	98.92	99.31	95.06	34.18
19276	98.86	99.28	98.75	98.87	99.29	95.03	33.85
50045	98.95	99.33	98.79	98.95	99.35	95.06	34.55
50179	98.88	99.28	98.77	98.90	99.30	95.05	34.27
50248	98.95	99.39	98.82	98.97	99.36	95.13	35.09
53221	99.01	99.43	98.87	99.03	99.41	95.16	35.10
54902	98.97	99.38	98.84	99.00	99.37	95.13	34.83
55546	98.72	99.13	98.78	98.93	99.19	95.10	34.88
55753	98.94	99.35	98.83	98.96	99.34	95.08	34.47
55988	98.98	99.40	98.85	99.00	99.38	95.14	35.22
55989	98.80	99.22	98.87	99.01	99.28	95.15	35.14
56828	99.08	99.35	98.69	98.85	99.21	95.21	35.48
57052	98.86	99.28	98.72	98.86	99.29	95.01	33.95
58385	98.83	99.27	98.73	98.86	99.27	95.05	34.57
60053	98.77	99.19	98.85	98.98	99.22	95.13	35.25
62796	98.92	99.32	98.80	98.92	99.33	95.07	35.01
62806	98.90	99.31	98.78	98.89	99.31	95.06	35.08
62957	98.93	99.34	98.78	98.94	99.33	95.10	35.07
63113	98.97	99.39	98.84	98.98	99.38	95.12	35.13
63670	98.89	99.32	98.79	98.91	99.30	95.03	34.79
63878	98.93	99.34	98.81	98.92	99.34	95.06	34.87

64165	98.90	99.32	98.78	98.91	99.30	95.08	34.78
64334	98.73	99.14	98.78	98.93	99.18	95.08	35.15
64712	98.90	99.32	98.77	98.91	99.31	95.07	34.80
65165	98.69	99.09	98.77	98.91	99.16	95.06	34.92
66591	98.50	98.89	98.57	98.70	98.95	94.89	34.50
68995	98.79	99.19	98.85	99.01	99.21	95.14	34.54
69094	99.06	99.50	98.93	99.09	99.45	95.23	35.38
73787	99.08	99.52	98.93	99.09	99.48	95.19	35.28
74856	99.00	99.44	98.87	99.03	99.42	95.17	34.92
78501	99.00	99.43	98.86	98.99	99.39	95.16	34.83
78932	98.98	99.42	98.84	99.04	99.40	95.18	34.93
79031	98.84	99.28	98.91	99.08	99.28	95.21	34.83
MT-0080	98.77	99.23	98.69	98.81	99.23	95.02	34.00
MT-0712	98.90	99.32	98.78	98.92	99.31	95.09	34.28
MT-0718	98.85	99.31	98.76	98.89	99.28	95.07	34.30
MT-0721	98.86	99.28	98.75	98.86	99.28	95.03	34.09
MT-0751	98.66	99.09	98.79	98.88	99.13	95.09	34.95
MT-0972	98.90	99.33	98.81	98.95	99.34	95.13	35.16
MT-1103	98.89	99.31	98.80	98.90	99.31	95.10	34.67
MT-1128	98.89	99.31	98.79	98.93	99.32	95.09	34.30
MT-1167	98.53	98.95	98.65	98.76	99.02	94.97	33.34
MT-1206	98.87	99.31	98.74	98.90	99.29	95.05	34.15
MT-1212	98.94	99.37	98.82	98.97	99.36	95.15	34.81
MT-1247	98.90	99.18	98.50	98.64	99.05	95.06	34.18
MT-13-1408	99.03	99.31	98.68	98.81	99.16	95.19	34.80
MT-13-1711	99.06	99.32	98.66	98.80	99.19	95.17	34.71
MT-13-1712	98.97	99.21	98.61	98.71	99.10	95.11	34.37
MT-13-1753	98.73	99.17	98.82	98.98	99.21	95.12	34.80
MT-13-1828	99.07	99.33	98.67	98.80	99.19	95.17	34.81
MT-13-1835	99.03	99.29	98.65	98.78	99.15	95.15	34.35
MT-13-1892	98.99	99.26	98.59	98.74	99.13	95.11	34.62
MT-13-2012	98.98	99.20	98.59	98.73	99.09	95.12	35.06
MT-13-2334	98.94	99.38	98.81	98.98	99.35	95.10	34.45
MT-13-2384	98.72	99.15	98.80	98.94	99.19	95.09	34.63
MT-13-2690	99.01	99.44	98.88	99.03	99.41	95.15	34.76
MT-13-2761	98.93	99.38	98.80	98.95	99.35	95.07	34.13
MT-13-3209	98.74	99.17	98.84	98.98	99.21	95.12	34.68
MT-13-848	98.96	99.24	98.61	98.71	99.11	95.16	34.86
MT-131	98.56	98.96	98.66	98.78	99.04	94.96	33.19
MT-1336	98.81	99.24	98.72	98.83	99.26	95.03	33.84
MT-1345	98.67	99.12	98.75	98.88	99.17	95.03	34.17
MT-1393	98.87	99.30	98.78	98.88	99.29	95.05	34.25

MT-140	98.86	99.28	98.73	98.87	99.28	95.05	33.98
MT-1403	98.80	99.23	98.69	98.82	99.24	95.02	34.30
MT-1466	98.88	99.34	98.77	98.94	99.32	95.09	34.47
MT-1499	98.80	99.23	98.68	98.83	99.23	95.04	34.43
MT-1549	98.89	99.32	98.77	98.90	99.32	95.10	34.36
MT-1605	98.84	99.29	98.75	98.90	99.31	95.08	34.38
MT-1684	98.88	99.30	98.77	98.90	99.31	95.12	35.29
MT-1799	98.87	99.29	98.75	98.88	99.28	95.08	34.71
MT-1838	98.89	99.31	98.78	98.92	99.31	95.11	34.87
MT-1971	98.89	99.32	98.76	98.91	99.32	95.05	34.36
MT-2151	98.86	99.28	98.76	98.88	99.29	95.07	34.42
MT-2174	98.83	99.25	98.73	98.87	99.27	95.02	33.81
MT-2175	98.80	99.20	98.70	98.83	99.24	95.03	33.89
MT-2178	98.67	99.08	98.75	98.86	99.12	95.05	34.27
MT-2184	98.86	99.29	98.77	98.89	99.30	95.08	34.41
MT-2224	98.79	99.23	98.68	98.80	99.24	95.00	34.03
MT-2356	98.91	99.32	98.80	98.90	99.32	95.10	34.67
MT-2465	98.90	99.33	98.81	98.93	99.33	95.11	34.64
MT-2473	98.82	99.26	98.76	98.87	99.26	95.09	34.67
MT-2474	98.76	99.19	98.69	98.79	99.22	94.98	33.57
MT-2538	98.89	99.30	98.77	98.89	99.30	95.08	34.64
MT-2665	98.82	99.21	98.71	98.83	99.23	95.02	33.78
MT-2667	98.84	99.27	98.73	98.85	99.28	95.08	34.42
MT-2706	98.85	99.28	98.74	98.87	99.28	95.03	34.00
MT-2720	98.85	99.31	98.77	98.90	99.30	95.10	34.70
MT-2762	98.83	99.26	98.74	98.85	99.26	95.04	34.40
MT-2768	98.91	99.32	98.77	98.92	99.32	95.09	34.21
MT-2769	98.86	99.28	98.77	98.87	99.28	95.07	34.54
MT-2771	98.82	99.24	98.73	98.85	99.28	95.08	34.61
MT-2792	98.57	98.96	98.64	98.76	99.02	94.94	34.33
MT-2800	98.90	99.33	98.81	98.94	99.32	95.12	34.64
MT-289	98.93	99.38	98.83	98.97	99.38	95.13	34.69
MT-2905	98.48	98.89	98.59	98.69	98.96	94.87	34.02
MT-2910	98.55	98.96	98.64	98.77	99.01	94.93	34.53
MT-2931	98.61	99.04	98.67	98.81	99.08	95.00	35.18
MT-3000	98.64	99.04	98.65	98.80	99.09	94.94	34.28
MT-3004	98.53	98.94	98.62	98.75	98.99	94.96	34.72
MT-3074	98.79	99.22	98.70	98.80	99.22	95.01	33.66
MT-3173	98.66	99.08	98.56	98.65	99.08	94.90	33.11
MT-3194	98.84	99.27	98.74	98.88	99.26	95.11	34.61
MT-3239	98.61	99.02	98.67	98.80	99.10	95.02	34.17
MT-3255	98.71	99.13	98.62	98.73	99.15	94.95	33.11

MT-3271	98.79	99.22	98.72	98.80	99.22	94.97	33.48
MT-3281	98.62	99.04	98.72	98.85	99.11	95.05	34.42
MT-3296	98.77	99.17	98.82	98.97	99.21	95.10	34.71
MT-3341	98.76	99.22	98.67	98.79	99.21	95.00	33.72
MT-3673	98.89	99.30	98.76	98.90	99.31	95.07	34.12
MT-3683	98.81	99.25	98.72	98.85	99.24	95.06	34.34
MT-3787	98.82	99.25	98.72	98.84	99.25	95.04	33.92
MT-389	98.82	99.26	98.72	98.87	99.26	95.06	34.04
MT-393	98.68	99.09	98.75	98.87	99.15	95.07	34.66
MT-398	98.78	99.18	98.64	98.75	99.20	94.97	33.36
MT-405	98.84	99.27	98.75	98.86	99.27	95.07	34.22
MT-4067	98.81	99.24	98.72	98.84	99.26	95.03	34.12
MT-4137	98.68	99.11	98.77	98.91	99.16	95.07	34.44
MT-4166	98.99	99.41	98.87	99.00	99.41	95.16	35.33
MT-4230	98.79	99.20	98.85	99.00	99.24	95.14	35.10
MT-441	98.95	99.36	98.80	98.95	99.37	95.09	34.49
MT-452	98.94	99.40	98.84	98.98	99.39	95.14	34.63
MT-467	98.80	99.21	98.69	98.83	99.23	94.99	33.61
MT-4683	98.88	99.30	98.75	98.89	99.29	95.07	34.42
MT-4846	98.47	98.88	98.58	98.67	98.97	94.93	32.92
MT-4854	98.82	99.25	98.72	98.86	99.25	95.06	33.95
MT-4884	98.63	99.07	98.72	98.85	99.12	95.05	34.33
MT-4942	98.77	99.20	98.68	98.79	99.22	95.00	33.83
MT-504	98.98	99.44	98.87	99.00	99.41	95.15	34.94
MT-5195	98.95	99.38	98.83	98.98	99.38	95.12	34.82
MT-5337	98.71	99.14	98.64	98.73	99.15	94.97	33.36
MT-5373	98.93	99.34	98.80	98.92	99.33	95.13	35.07
MT-5383	98.85	99.28	98.76	98.88	99.28	95.05	34.52
MT-5447	98.71	99.09	98.77	98.93	99.16	95.09	34.95
MT-5531	98.70	99.13	98.64	98.73	99.17	94.95	33.32
MT-5543	98.65	99.08	98.74	98.87	99.15	95.05	34.39
MT-567	98.82	99.23	98.73	98.84	99.25	95.06	34.39
MT-5870	98.85	99.27	98.72	98.85	99.27	95.05	34.16
MT-5983	98.84	99.26	98.77	98.84	99.26	95.11	34.65
MT-6084	98.82	99.25	98.73	98.84	99.24	95.07	34.25
MT-6205	98.62	99.06	98.72	98.86	99.12	95.05	34.61
MT-6218	98.90	99.34	98.81	98.93	99.35	95.11	34.91
MT-6226	98.73	99.17	98.65	98.73	99.18	94.96	33.35
MT-6429	98.90	99.33	98.80	98.94	99.35	95.13	35.16
MT-661	98.90	99.33	98.78	98.90	99.31	95.07	34.33
MT-692	98.92	99.34	98.90	98.97	99.36	95.16	38.20
MT-853	98.83	99.24	98.71	98.83	99.24	95.01	34.34

MT-877	98.92	99.36	98.78	98.92	99.34	95.08	34.43

TABLE S3 Depth of coverage for each isolate

	Average depth of coverage by reference genome						
Isolate	*M. tuberculosis*, lineage 4 H37Rv	*M. tuberculosis*, lineage 4 CDC1551	*M. tuberculosis*, lineage 2 CCDC5079	*M. africanum* GN041182	*M. bovis* AF2122/97	*M. canettii* CIPT 140010059	*M. kansasii* ATCC 12478
9965	68.99	69.26	68.98	68.90	69.60	66.41	15.03
10155	47.63	47.81	47.61	47.55	48.06	45.82	10.91
10223	79.56	79.88	79.70	79.62	80.27	76.70	17.65
11011	49.04	49.22	49.03	48.95	49.45	47.18	11.26
11234	42.84	43.00	42.83	42.76	43.20	41.22	9.63
14069	51.64	51.82	51.62	51.54	52.08	49.67	11.75
14508	79.36	79.69	79.53	79.45	80.03	76.58	17.60
15613	73.42	73.73	73.40	73.34	74.11	70.63	15.77
16490	57.99	58.23	57.98	57.92	58.48	55.78	12.65
16493	58.34	58.56	58.33	58.22	58.84	56.15	13.12
18421	43.40	43.58	43.39	43.34	43.79	41.76	9.84
18422	65.68	65.92	65.68	65.59	66.27	63.20	14.72
18747	51.66	51.87	51.65	51.59	52.19	49.70	11.76
18988	49.43	49.63	49.42	49.35	49.95	47.53	11.18
19057	49.32	49.50	49.31	49.25	49.81	47.47	11.27
19276	40.47	40.63	40.45	40.41	40.90	38.92	9.22
50045	55.85	56.07	55.83	55.79	56.34	53.70	12.16
50179	53.03	53.24	53.03	52.96	53.56	51.05	12.22
50248	80.81	81.13	80.80	80.72	81.54	77.75	17.40
53221	75.43	75.72	75.39	75.35	76.07	72.58	17.16
54902	59.53	59.77	59.51	59.45	60.06	57.26	13.31
55546	89.11	89.47	89.25	89.21	89.90	85.92	19.69
55753	52.70	52.90	52.71	52.64	53.14	50.71	11.77
55988	98.92	99.31	98.89	98.82	99.75	95.26	22.72
55989	89.10	89.44	89.21	89.18	89.86	85.93	20.67
56828	111.65	111.89	111.37	111.23	112.33	107.49	25.85
57052	40.81	40.98	40.80	40.76	41.21	39.26	9.31
58385	74.28	74.57	74.25	74.19	74.94	71.46	15.76
60053	84.05	84.37	84.16	84.14	84.76	81.10	19.00
62796	77.14	77.43	77.11	77.06	77.79	74.27	17.23
62806	88.45	88.77	88.39	88.35	89.21	85.14	20.01
62957	74.01	74.30	73.99	73.97	74.67	71.25	16.20
63113	79.20	79.51	79.18	79.14	79.92	76.24	17.98
63670	68.14	68.41	68.12	68.07	68.79	65.62	15.57
63878	73.12	73.40	73.10	73.05	73.75	70.41	16.69

64165	65.98	66.23	65.95	65.92	66.59	63.53	14.78
64334	83.68	84.01	83.78	83.80	84.41	80.70	18.81
64712	70.07	70.33	70.04	70.00	70.66	67.48	15.68
65165	75.17	75.47	75.26	75.27	75.82	72.57	17.04
66591	58.02	58.25	58.10	58.09	58.51	56.01	13.18
68995	54.47	54.69	54.56	54.54	54.86	52.49	12.26
69094	89.53	89.86	89.52	89.42	90.22	86.12	19.87
73787	71.92	72.18	71.88	71.83	72.48	69.16	16.54
74856	73.43	73.73	73.42	73.35	74.03	70.61	15.85
78501	61.82	62.03	61.80	61.75	62.32	59.46	13.70
78932	68.42	68.68	68.40	68.35	68.97	65.82	15.11
79031	62.64	62.86	62.70	62.70	63.17	60.40	14.48
MT-0080	58.44	58.65	58.45	58.36	58.93	56.25	12.85
MT-0712	53.75	53.95	53.74	53.68	54.21	51.73	11.95
MT-0718	54.04	54.24	54.04	53.97	54.54	52.03	11.81
MT-0721	48.33	48.50	48.34	48.26	48.77	46.55	10.53
MT-0751	48.42	48.59	48.50	48.45	48.84	46.70	10.74
MT-0972	88.00	88.33	87.98	87.88	88.75	84.72	20.13
MT-1103	70.91	71.17	70.89	70.79	71.50	68.24	15.79
MT-1128	48.84	49.01	48.83	48.75	49.27	47.03	11.14
MT-1167	43.48	43.64	43.55	43.50	43.89	41.85	8.69
MT-1206	51.48	51.65	51.47	51.40	51.92	49.52	11.14
MT-1212	73.74	74.01	73.72	73.64	74.38	70.96	16.05
MT-1247	58.07	58.19	57.94	57.88	58.43	55.93	13.00
MT-13-1408	78.78	78.94	78.59	78.50	79.25	75.85	17.81
MT-13-1711	60.34	60.45	60.19	60.10	60.70	58.10	13.73
MT-13-1712	60.17	60.29	60.06	59.95	60.54	57.93	13.00
MT-13-1753	70.94	71.21	71.04	70.97	71.51	68.38	15.95
MT-13-1828	69.56	69.70	69.39	69.30	70.01	66.97	15.58
MT-13-1835	48.67	48.78	48.56	48.51	48.97	46.89	11.00
MT-13-1892	61.89	62.01	61.73	61.64	62.25	59.59	13.98
MT-13-2012	108.73	108.98	108.46	108.33	109.37	104.68	23.30
MT-13-2334	56.86	57.07	56.85	56.79	57.37	54.71	12.44
MT-13-2384	63.19	63.42	63.28	63.22	63.72	60.92	14.13
MT-13-2690	66.11	66.35	66.10	66.03	66.68	63.62	14.63
MT-13-2761	49.01	49.19	49.01	48.95	49.45	47.18	10.79
MT-13-3209	68.42	68.66	68.51	68.44	69.00	65.96	15.26
MT-13-848	92.69	92.91	92.46	92.38	93.27	89.27	20.29
MT-131	39.42	39.57	39.48	39.41	39.82	37.97	8.11
MT-1336	46.87	47.02	46.88	46.76	47.29	45.09	10.23
MT-1345	48.42	48.58	48.49	48.42	48.84	46.72	11.06
MT-1393	52.67	52.84	52.65	52.55	53.12	50.69	11.88

MT-140	42.75	42.89	42.75	42.66	43.12	41.15	9.77
MT-1403	74.69	74.97	74.69	74.59	75.32	71.89	16.09
MT-1466	53.41	53.60	53.41	53.33	53.90	51.41	11.66
MT-1499	71.15	71.41	71.17	71.04	71.75	68.48	15.53
MT-1549	53.16	53.35	53.17	53.06	53.62	51.13	11.95
MT-1605	55.26	55.47	55.24	55.18	55.75	53.17	12.48
MT-1684	123.56	123.99	123.54	123.34	124.59	118.86	27.31
MT-1799	80.97	81.26	80.95	80.86	81.66	77.94	18.15
MT-1838	81.86	82.16	81.87	81.72	82.56	78.82	18.60
MT-1971	49.45	49.60	49.45	49.33	49.86	47.59	11.30
MT-2151	62.35	62.59	62.36	62.28	62.91	60.03	14.07
MT-2174	38.15	38.27	38.14	38.06	38.47	36.69	8.26
MT-2175	43.85	43.98	43.84	43.75	44.22	42.16	9.63
MT-2178	50.66	50.81	50.74	50.65	51.12	48.83	11.32
MT-2184	56.40	56.58	56.39	56.25	56.88	54.26	12.58
MT-2224	51.33	51.49	51.32	51.20	51.79	49.36	10.92
MT-2356	62.91	63.10	62.91	62.78	63.46	60.54	14.20
MT-2465	62.19	62.41	62.18	62.06	62.73	59.84	14.19
MT-2473	80.14	80.43	80.13	80.03	80.84	77.13	17.72
MT-2474	49.12	49.30	49.11	49.04	49.59	47.27	10.27
MT-2538	60.33	60.52	60.32	60.19	60.87	58.04	13.60
MT-2665	47.72	47.89	47.71	47.64	48.13	45.92	10.54
MT-2667	61.75	61.98	61.76	61.68	62.33	59.48	13.64
MT-2706	42.99	43.13	43.00	42.90	43.36	41.35	9.53
MT-2720	67.83	68.05	67.81	67.69	68.40	65.29	15.47
MT-2762	56.61	56.81	56.61	56.51	57.10	54.52	12.81
MT-2768	46.93	47.08	46.92	46.84	47.35	45.19	10.51
MT-2769	63.40	63.60	63.40	63.28	63.95	61.05	14.32
MT-2771	80.10	80.38	80.10	79.98	80.80	77.08	17.65
MT-2792	60.65	60.86	60.75	60.65	61.17	58.52	13.46
MT-2800	64.85	65.08	64.85	64.74	65.37	62.44	14.83
MT-289	68.07	68.34	68.07	67.98	68.64	65.52	15.43
MT-2905	56.03	56.25	56.12	56.08	56.49	54.05	12.51
MT-2910	69.75	70.00	69.87	69.81	70.38	67.36	15.90
MT-2931	104.96	105.40	105.16	105.09	105.93	101.32	23.75
MT-3000	53.55	53.73	53.64	53.56	54.01	51.64	11.88
MT-3004	75.17	75.41	75.30	75.17	75.78	72.50	16.71
MT-3074	49.85	50.00	49.85	49.74	50.29	47.97	10.39
MT-3173	42.06	42.23	42.05	42.04	42.45	40.47	8.63
MT-3194	72.91	73.18	72.92	72.82	73.55	70.16	15.96
MT-3239	71.75	72.00	71.87	71.78	72.36	69.21	15.68
MT-3255	44.38	44.54	44.38	44.31	44.76	42.71	9.55

MT-3271	47.46	47.63	47.47	47.39	47.86	45.61	10.20
MT-3281	80.10	80.38	80.24	80.12	80.78	77.25	17.26
MT-3296	61.10	61.29	61.19	61.09	61.64	58.91	13.89
MT-3341	46.84	47.01	46.84	46.75	47.26	45.07	10.18
MT-3673	48.90	49.05	48.89	48.80	49.33	47.03	10.74
MT-3683	70.56	70.82	70.56	70.45	71.17	67.93	15.47
MT-3787	45.52	45.68	45.51	45.43	45.92	43.80	10.15
MT-389	48.13	39.43	48.13	48.02	48.53	46.30	10.67
MT-393	67.52	67.74	67.62	67.54	68.10	65.15	15.46
MT-398	39.29	39.43	39.28	39.23	39.62	37.83	8.67
MT-405	53.99	54.17	53.99	53.87	54.46	51.95	11.96
MT-4067	59.17	59.37	59.16	59.07	59.65	56.93	13.02
MT-4137	53.08	53.24	53.17	53.09	53.55	51.20	11.95
MT-4166	96.86	97.22	96.82	96.73	97.70	93.22	21.74
MT-4230	91.26	91.62	91.40	91.35	92.04	88.05	20.36
MT-441	54.34	54.51	54.33	54.21	54.80	52.31	12.39
MT-452	55.69	55.89	55.70	55.59	56.21	53.61	12.36
MT-467	47.90	48.10	47.93	47.87	48.35	46.11	10.31
MT-4683	67.69	67.91	67.69	67.58	68.28	65.13	14.95
MT-4846	53.03	53.26	53.13	53.10	53.61	51.08	10.29
MT-4854	53.10	53.28	53.10	53.00	53.54	51.09	11.75
MT-4884	65.72	65.95	65.80	65.75	66.25	63.37	14.69
MT-4942	60.97	61.19	60.98	60.87	61.50	58.68	13.13
MT-504	67.54	67.77	67.53	67.42	68.10	64.99	15.88
MT-5195	76.29	76.57	76.29	76.17	76.94	73.44	17.60
MT-5337	50.84	51.03	50.85	50.77	51.32	48.94	10.64
MT-5373	96.44	96.80	96.42	96.29	97.25	92.83	21.46
MT-5383	61.66	61.90	61.67	61.61	62.23	59.31	12.84
MT-5447	55.46	55.69	55.56	55.52	56.05	53.44	11.78
MT-5531	56.13	56.39	56.15	56.10	56.73	54.00	11.29
MT-5543	58.77	58.96	58.86	58.78	59.28	56.70	12.82
MT-567	77.89	78.17	77.89	77.76	78.54	74.96	16.78
MT-5870	50.87	51.04	50.86	50.77	51.31	48.96	11.11
MT-5983	76.06	76.39	76.07	76.00	76.91	73.13	15.39
MT-6084	69.98	70.23	69.98	69.86	70.58	67.36	15.11
MT-6205	73.78	74.10	73.89	73.87	74.54	71.13	15.46
MT-6218	74.27	74.56	74.27	74.17	74.91	71.50	17.00
MT-6226	52.55	52.81	52.57	52.54	53.15	50.52	10.43
MT-6429	88.53	88.87	88.52	88.39	89.33	85.24	20.10
MT-661	49.89	50.05	49.88	49.79	50.32	47.99	11.08
MT-692	66.84	67.09	66.84	66.74	67.43	64.32	14.61
MT-853	59.55	59.75	59.56	59.45	60.08	57.31	13.03

MT-877	51.41	51.57	51.41	51.31	51.86	49.49	11.43

TABLE S4 Clusters and colour scheme used in Figures

			Changes in clustering compared to those identified using CDC1551, by reference genome					
Isolate	Cluster according to *M. tuberculosis* lineage 4, CDC1551	Colours used in all figures based on the CDC1551 clusters	*M. tuberculosis* H37Rv	*M. tuberculosis* CCDC5079	*M. africanum*	*M. bovis*	*M. canettii*	*M. kansasii*
56828	Mn		No change	No change	No change	No change	No change	Clustering lost
MT-1247	Mn		No change	No change	No change	No change	No change	Clustering lost
MT-13-1408	Mn		No change	No change	No change	No change	No change	Clustering lost
MT-13-1711	Mn		No change	No change	No change	No change	No change	Clustering lost
MT-13-1712	Mn		No change	No change	No change	No change	No change	Clustering lost
MT-13-1828	Mn		No change	No change	No change	No change	No change	Clustering lost
MT-13-1835	Mn		No change	No change	No change	No change	No change	Clustering lost
MT-13-1892	Mn		No change	No change	No change	No change	No change	Clustering lost
MT-13-2012	Mn		No change	No change	No change	No change	No change	Clustering lost
MT-13-848	Mn		No change	No change	No change	No change	No change	Clustering lost
55753	Mj-VI		No change	No change	No change	No change	No change	No change
55988	Mj-VI		No change	No change	No change	No change	No change	No change
10155	Mj-I		No change	No change	No change	No change	No change	No change
11011	Mj-I		No change	No change	No change	No change	No change	No change
11234	Mj-I		No change	No change	No change	No change	No change	No change
14069	Mj-I		No change	No change	No change	No change	No change	No change
16493	Mj-I		No change	No change	No change	No change	No change	No change
18421	Mj-I		No change	No change	No change	No change	No change	No change
18422	Mj-I		No change	No change	No change	No change	No change	No change
18747	Mj-II		No change	No change	No change	No change	No change	Clustering lost
18988	Mj-II		No change	No change	No change	No change	No change	Clustering lost
19057	Mj-II		No change	No change	No change	No change	No change	Clustering lost

19276	Mj-II		No change	No change	No change	No change	No change	Clustering lost
50179	Mj-II		No change	No change	No change	No change	No change	Clustering lost
57052	Mj-II		No change	No change	No change	No change	No change	Clustering lost
63670	Mj-II		No change	No change	No change	No change	No change	Clustering lost
64165	Mj-II		No change	No change	No change	No change	No change	Clustering lost
54902	Mj-IV.a		No change	No change	No change	No change	No change	Clustering lost
9965	Mj-IV.a		No change	No change	No change	No change	No change	Clustering lost
16490	Mj-IV.b		No change	No change	No change	No change	No change	Clustering lost
50045	Mj-IV.b		No change	No change	No change	No change	No change	Clustering lost
50248	Mj-IV.b		No change	No change	No change	No change	No change	Clustering lost
53221	Mj-IV.b		No change	No change	No change	No change	No change	Clustering lost
62806	Mj-IV.b		No change	No change	No change	No change	No change	Clustering lost
62957	Mj-IV.b		No change	No change	No change	No change	No change	Clustering lost
63113	Mj-IV.b		No change	No change	No change	No change	No change	Clustering lost
64712	Mj-IV.b		No change	No change	No change	No change	No change	Clustering lost
MT-13-2334	Mj-IV.b		No change	No change	No change	No change	No change	Clustering lost
MT-13-2690	Mj-IV.b		No change	No change	No change	No change	No change	Clustering lost
MT-13-2761	Mj-IV.b		No change	No change	No change	No change	No change	Clustering lost
MT-1403	Mj-IV.b		No change	No change	No change	No change	No change	Clustering lost
MT-1799	Mj-IV.b		No change	No change	No change	No change	No change	Clustering lost
MT-398	Mj-IV.b		No change	No change	No change	No change	No change	Clustering lost
MT-4067	Mj-IV.b		No change	No change	No change	No change	No change	Clustering lost
MT-4683	Mj-IV.b		No change	No change	No change	No change	No change	Clustering lost
MT-5373	Mj-IV.b		No change	No change	No change	No change	No change	Clustering lost
MT-5870	Mj-IV.b		No change	No change	No change	No change	No change	Clustering lost
62796	Mj-IV.c		No change	No change	No change	No change	No change	Clustering lost
63878	Mj-IV.c		No change	No change	No change	No change	No change	Clustering lost
MT-140	Mj-IV.c		No change	No change	No change	No change	No change	Clustering lost
MT-1499	Mj-IV.c		No change	No change	No change	No change	No change	Clustering lost
MT-1971	Mj-IV.c		No change	No change	No change	No change	No change	Clustering lost

MT-2224	Mj-IV.c		No change	No change	No change	No change	No change	Clustering lost
MT-2768	Mj-IV.c		No change	No change	No change	No change	No change	Clustering lost
MT-441	Mj-IV.c		No change	No change	No change	No change	No change	Clustering lost
MT-452	Mj-IV.c		No change	No change	No change	No change	No change	Clustering lost
MT-661	Mj-IV.c		No change	No change	No change	No change	No change	Clustering lost
MT-692	Mj-IV.c		No change	No change	No change	No change	No change	Clustering lost
MT-853	Mj-IV.c		No change	No change	No change	No change	No change	Clustering lost
MT-877	Mj-IV.c		No change	No change	No change	No change	No change	Clustering lost
58385	Not clustered		No change	No change	No change	No change	No change	No change
10223	Mj-V.a		No change	No change	No change	No change	No change	Clustering lost
55546	Mj-V.a		No change	No change	No change	No change	No change	Clustering lost
55989	Mj-V.a		No change	No change	No change	No change	No change	Clustering lost
60053	Mj-V.a		No change	No change	No change	No change	No change	Clustering lost
64334	Mj-V.a		No change	No change	No change	No change	No change	Clustering lost
65165	Mj-V.a		No change	No change	No change	No change	No change	Clustering lost
68995	Mj-V.a		No change	No change	No change	No change	No change	Clustering lost
MT-1167	Mj-V.a		No change	No change	No change	No change	No change	Clustering lost
MT-13-1753	Mj-V.a		No change	No change	No change	No change	No change	Clustering lost
MT-1345	Mj-V.a		No change	No change	No change	No change	No change	Clustering lost
MT-3296	Mj-V.a		No change	No change	No change	No change	No change	Clustering lost
MT-393	Mj-V.a		No change	No change	No change	No change	No change	Clustering lost
MT-4137	Mj-V.a		No change	No change	No change	No change	No change	Clustering lost
MT-4884	Mj-V.a		No change	No change	No change	No change	No change	Clustering lost
MT-5543	Mj-V.a		No change	No change	No change	No change	No change	Clustering lost
MT-751	Mj-V.a		No change	No change	No change	No change	No change	Clustering lost
14508	Mj-V.a		No change	No change	No change	No change	Mj-V.c	Clustering lost
79031	Mj-V.b		No change	No change	No change	No change	No change	Clustering lost
MT-13-2384	Mj-V.b		No change	No change	No change	No change	No change	Clustering lost
MT-13-3209	Mj-V.b		No change	No change	No change	No change	No change	Clustering lost
MT-131	Mj-V.b		No change	No change	No change	No change	No change	Clustering lost

MT-2178	Mj-V.b		No change	No change	No change	No change	No change	Clustering lost
MT-3281	Mj-V.b		No change	No change	No change	No change	No change	Clustering lost
MT-4230	Mj-V.b		No change	No change	No change	No change	No change	Clustering lost
MT-4846	Mj-V.b		No change	No change	No change	No change	No change	Clustering lost
MT-5447	Mj-V.b		No change	No change	No change	No change	No change	Clustering lost
MT-6205	Mj-V.b		No change	No change	No change	No change	No change	Clustering lost
66591	Mj-V.c		No change	No change	No change	No change	No change	Clustering lost
MT-2792	Mj-V.c		No change	No change	No change	No change	No change	Clustering lost
MT-2905	Mj-V.c		No change	No change	No change	No change	No change	Clustering lost
MT-2910	Mj-V.c		No change	No change	No change	No change	No change	Clustering lost
MT-2931	Mj-V.c		No change	No change	No change	No change	No change	Clustering lost
MT-3000	Mj-V.c		No change	No change	No change	No change	No change	Clustering lost
MT-3004	Mj-V.c		No change	No change	No change	No change	No change	Clustering lost
MT-3239	Mj-V.c		No change	No change	No change	No change	No change	Clustering lost
15613	Mj-V.d		No change	No change	No change	No change	No change	Clustering lost
MT-2538	Mj-V.d		No change	No change	No change	No change	No change	Clustering lost
74856	Mj-III.a		No change	No change	No change	No change	No change	Clustering lost
78932	Mj-III.a		No change	No change	No change	No change	No change	Clustering lost
MT-1393	Mj-III.a		No change	No change	No change	No change	No change	Clustering lost
MT-1549	Mj-III.a		No change	No change	No change	No change	No change	Clustering lost
MT-1605	Mj-III.a		No change	No change	No change	No change	No change	Clustering lost
MT-2175	Mj-III.a		No change	No change	No change	No change	No change	Clustering lost
MT-2706	Mj-III.a		No change	No change	No change	No change	No change	Clustering lost
MT-2720	Mj-III.a		No change	No change	No change	No change	No change	Clustering lost
MT-2771	Mj-III.a		No change	No change	No change	No change	No change	Clustering lost
MT-2800	Mj-III.a		No change	No change	No change	No change	No change	Clustering lost
MT-3074	Mj-III.a		No change	No change	No change	No change	No change	Clustering lost
MT-3255	Mj-III.a		No change	No change	No change	No change	No change	Clustering lost
MT-3341	Mj-III.a		No change	No change	No change	No change	No change	Clustering lost
MT-3673	Mj-III.a		No change	No change	No change	No change	No change	Clustering lost

MT-4166	Mj-III.a		No change	No change	No change	No change	No change	Clustering lost
MT-4942	Mj-III.a		No change	No change	No change	No change	No change	Clustering lost
MT-5195	Mj-III.a		No change	No change	No change	No change	No change	Clustering lost
MT-5337	Mj-III.a		No change	No change	No change	No change	No change	Clustering lost
MT-5383	Mj-III.a		No change	No change	No change	No change	No change	Clustering lost
MT-5488	Mj-III.a		No change	No change	No change	No change	No change	Clustering lost
MT-5531	Mj-III.a		No change	No change	No change	No change	No change	Clustering lost
MT-5983	Mj-III.a		No change	No change	No change	No change	No change	Clustering lost
MT-721	Mj-III.a		No change	No change	No change	No change	No change	Clustering lost
73787	Mj-III.b		No change	No change	No change	No change	No change	Clustering lost
MT-1206	Mj-III.b		No change	No change	No change	No change	No change	Clustering lost
MT-1212	Mj-III.b		No change	No change	No change	No change	No change	Clustering lost
MT-1336	Mj-III.b		No change	No change	No change	No change	No change	Clustering lost
MT-1466	Mj-III.b		No change	No change	No change	No change	No change	Clustering lost
MT-1684	Mj-III.b		No change	No change	No change	No change	No change	Clustering lost
MT-2184	Mj-III.b		No change	No change	No change	No change	No change	Clustering lost
MT-2356	Mj-III.b		No change	No change	No change	No change	No change	Clustering lost
MT-2465	Mj-III.b		No change	No change	No change	No change	No change	Clustering lost
MT-2473	Mj-III.b		No change	No change	No change	No change	No change	Clustering lost
MT-2474	Mj-III.b		No change	No change	No change	No change	No change	Clustering lost
MT-2762	Mj-III.b		No change	No change	No change	No change	No change	Clustering lost
MT-3173	Mj-III.b		No change	No change	No change	No change	No change	Clustering lost
MT-3194	Mj-III.b		No change	No change	No change	No change	No change	Clustering lost
MT-3271	Mj-III.b		No change	No change	No change	No change	No change	Clustering lost
MT-3683	Mj-III.b		No change	No change	No change	No change	No change	Clustering lost
MT-4854	Mj-III.b		No change	No change	No change	No change	No change	Clustering lost
MT-504	Mj-III.b		No change	No change	No change	No change	No change	Clustering lost
MT-6084	Mj-III.b		No change	No change	No change	No change	No change	Clustering lost
MT-6429	Mj-III.b		No change	No change	No change	No change	No change	Clustering lost
MT-0080	Mj-III.c		No change	No change	No change	No change	No change	Clustering lost

MT-0712	Mj-III.c		No change	No change	No change	No change	No change	Clustering lost
MT-0718	Mj-III.c		No change	No change	No change	No change	No change	Clustering lost
MT-0972	Mj-III.c		No change	No change	No change	No change	No change	Clustering lost
MT-1103	Mj-III.c		No change	No change	No change	No change	No change	Clustering lost
MT-1128	Mj-III.c		No change	No change	No change	No change	No change	Clustering lost
MT-1838	Mj-III.c		No change	No change	No change	No change	No change	Clustering lost
MT-2151	Mj-III.c		No change	No change	No change	No change	No change	Clustering lost
MT-2174	Mj-III.c		No change	No change	No change	No change	No change	Clustering lost
MT-2665	Mj-III.c		No change	No change	No change	No change	No change	Clustering lost
MT-2667	Mj-III.c		No change	No change	No change	No change	No change	Clustering lost
MT-2769	Mj-III.c		No change	No change	No change	No change	No change	Clustering lost
MT-289	Mj-III.c		No change	No change	No change	No change	No change	Clustering lost
MT-3787	Mj-III.c		No change	No change	No change	No change	No change	Clustering lost
MT-389	Mj-III.c		No change	No change	No change	No change	No change	Clustering lost
MT-405	Mj-III.c		No change	No change	No change	No change	No change	Clustering lost
MT-467	Mj-III.c		No change	No change	No change	No change	No change	Clustering lost
MT-567	Mj-III.c		No change	No change	No change	No change	No change	Clustering lost
MT-578	Mj-III.c		No change	No change	No change	No change	No change	Clustering lost
MT-6218	Mj-III.c		No change	No change	No change	No change	No change	Clustering lost
MT-6226	Mj-III.c		No change	No change	No change	No change	No change	Clustering lost

Supplemental references

1. **Lee RS, Radomski N, Proulx J-F, Levade I, Shapiro BJ, McIntosh F, Soualhine H, Menzies D, Behr MA.** 2015. Population genomics of *Mycobacterium tuberculosis* in the Inuit. Proc Natl Acad Sci U S A **112**(44):13609-13614.
2. **Tamura K.** 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. Mol Biol Evol **9**:678-687.
3. **Waddell PJ, Steel MA.** 1997. General time-reversible distances with unequal rates across sites: mixing Γ and inverse gaussian distributions with invariant sites. Mol Phylogenet Evol **8**:398-414.
4. **Richter M, Rosselló-Móra R.** 2009 Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A **106**(45):19126-19131.
5. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.** 1990. Basic local alignment search tool. J Mol Biol **215**:403-410.
6. **Delcher AL, Phillippy A, Carlton J, Salzberg SL**. 2002. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res **30**(11):2478-2483.
7. **Chan JZM, Halachev MR, Loman NJ, Constantinidou C, Pallen JM.** 2012. Defining bacterial species in the genomic era: insights from the genus *Acinetobacter*. BMC Microbiol **12**:302.
8. **Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM.** 2007. DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol **57**(1):81–91.

APPENDIX 6

Reprint of:

Lee RS and Behr MA. Does choice matter? The implications of whole genome sequencing in the control of tuberculosis. 2016. *Ther Adv Infect Dis*;3(2):47-62.

The implications of whole-genome sequencing in the control of tuberculosis

Robyn S. Lee and Marcel A. Behr

Ther Adv Infect Dis
(2010) 0(0) 1–16
DOI: 10.1177/2049936115624630

Abstract: The availability of whole-genome sequencing (WGS) as a tool for the diagnosis and clinical management of tuberculosis (TB) offers considerable promise in the fight against this stubborn epidemic. However, like other new technologies, the best application of WGS remains to be determined, for both conceptual and technical reasons. In this review, we consider the potential value of WGS in the clinical laboratory for the detection of *Mycobacterium tuberculosis* and the prediction of antibiotic resistance. We also discuss issues pertaining to data generation, interpretation and dissemination, given that WGS has to date been generally performed in research labs where results are not necessarily packaged in a clinician-friendly format. Although WGS is far more accessible now than it was in the past, the transition from a research tool to study TB into a clinical test to manage this disease may require further fine-tuning. Improvements will likely come through iterative efforts that involve both the laboratories ready to move TB into the genomic era and the front-line clinical/public health staff who will be interpreting the results to inform management decisions.

Keywords: clinical microbiology, diagnostics, drug resistance, *Mycobacterium tuberculosis*, whole-genome sequencing

Introduction

Owing to advances in technology and reductions in cost, whole-genome sequencing (WGS) has been transformed from a centralized service used by a select few to interrogate single genomes into a relatively decentralized lab technique used by many to detect and track infectious pathogens [Long *et al.* 2014; Price *et al.* 2014; SenGupta *et al.* 2014; Snitkin *et al.* 2012; Quick *et al.* 2014, 2015]. This transformation has not spared the mycobacterial genus, with a number of papers presenting its application to the characterization of *Mycobacterium tuberculosis* cases and outbreaks [Walker *et al.* 2013; Bryant *et al.* 2013; Gardy *et al.* 2011; Lee *et al.* 2015; Casali *et al.* 2014; Jamieson *et al.* 2014b; Stucki *et al.* 2015; Roetzer *et al.* 2013; Guerra-Assuncao *et al.* 2015]. In this review, we will consider the opportunities presented by WGS for clinical management of tuberculosis (TB) across two conceptual spaces: diagnosis (*M. tuberculosis* detection) and treatment (prediction of antibiotic resistance). We recognize that the greatest utility for WGS will likely lie in countries with the highest TB burdens; however, as WGS requires substantial financial and technical infrastructure, we have situated this review in the setting of a high-resource country where this method may be more imminently implemented.

A brief description of WGS

WGS begins at the bench, with the extraction and purification of genomic DNA. In very brief detail, this DNA is typically fragmented into shorter pieces, which are then sequenced in 'reads' of 100–500 base pairs (bps) for bench-top sequencers. There are a number of different sequencing platforms available [Loman *et al.* 2012a; Kwong *et al.* 2015; Heather and Chan, 2015]. The choice of platform depends largely on the question, which in turn is dictated by clinical needs. If the aim is to identify unknown organisms or to characterize a novel bacterium, one might prefer a sequencer that generates longer reads (such as the PacBio RS by Pacific Biosciences, Menlo Park, CA, USA), as such reads enable more accurate *de novo* assembly [Loman *et al.* 2012a]. If the goal is to speciate the microorganism, determine drug resistance or resolve transmission networks, sequencers producing short reads can be used. Among the benchtop sequencers generating

Correspondence to:
Marcel A. Behr, MD, MSc
McGill University Health Centre, Glen Site, 1001 Decarie Boulevard, Block E, Mail Drop Point #EM33211, Montréal, QC, H4A 3J1, Canada
marcel.behr@mcgill.ca

Robyn S. Lee, BScN, PhD candidate
McGill University, Department of Epidemiology, Biostatistics and Occupational Health, The Research Institute of the McGill University Health Centre and McGill International TB Centre, Montreal, QC, Canada

short read data, the most accurate platform currently available is the Illumina MiSeq (Illumina, San Diego, CA, USA) [Loman *et al.* 2012b] (though whether the difference in accuracy compared with another platform, the Ion Torrent PGM from ThermoFisher Scientific, Waltham, MA, USA, ultimately affects clinical inferences has been questioned [Harris *et al.* 2013]). In the analysis of such short read data, a reference-based approach is preferred [Loman *et al.* 2012a], wherein these reads are aligned ('mapped') to a reference genome. This is ideal for analysis of *M. tuberculosis*, given the absence of horizontal gene transfer in this species and the existence of complete, well-annotated reference genomes. Such a workflow for *M. tuberculosis* is illustrated in Figure 1.

With the Illumina MiSeq platform, short reads of up to 300 bps in length are produced. To identify the microorganism in question based on these reads, a variety of tools can be utilized. The Basic Local Alignment Search Tool (BLAST [Altschul *et al.* 1990]) compares reads with existing microbial DNA databases and uses an algorithm to identify the most likely microorganism. Other methods include classifying the microorganism based on how well reads align to conserved coding sequences within phyla or species ('clade-specific marker sequences' [Segata *et al.* 2012]) or *k*-mer-based approaches [Wood and Salzberg, 2014]. In the latter, reads are divided into segments of *k* bases in length (called '*k*-mers') that are compared with a database of known *k*-mer sequences from selected microorganisms. The best identification is determined as the microorganism with the highest proportion of matching *k*-mers.

Once reads have been assigned the identity '*M. tuberculosis*', they are subsequently mapped to the corresponding sequence on the reference genome to identify differences (i.e. variants) in the sample compared with this reference. There are several key considerations when performing such reference-based analyses. First, the choice of an appropriate reference genome is crucial; if the reference is too dissimilar from the isolate in question, large numbers of reads will not be mapped and these data (and all variation therein) will be ignored. Second, alignment to GC-rich repetitive regions can be difficult, as reads may map to more than one location, thereby producing inconclusive matches. Such regions include the PE-PPE family proteins, which comprise ~10% of the coding sequence of *M. tuberculosis* [Cole *et al.* 1998]. To reduce the risk of false-positive results, the PE-PPE regions and mobile elements are typically excluded from analyses [Comas *et al.* 2010; Roetzer *et al.* 2013]. Alternatively, one could perform targeted sequencing using a platform capable of generating longer reads that span repetitive regions. However, this would incur additional expense, as well as technical/bioinformatics requirements, and may not provide additional information of use for clinical applications.

Using a reference-based approach, single nucleotide polymorphisms (SNPs; i.e. a difference in a single base in the genome compared to the reference) and insertions/deletions (indels) present in the test isolate can be identified ('called') compared with the referent. This process, the quality control steps therein and the different tools used for identifying SNPs are reviewed in detail elsewhere in [Pabinger *et al.* 2014; Olson *et al.* 2015]. For the purposes of this work, we have focused on the utility of WGS for the clinician and, in particular, the use of these SNPs to predict drug resistance. In *M. tuberculosis* research, SNPs have also been used to extensively to delineate transmission networks, however, an in-depth discussion of this utility is beyond the scope of this review. The interested reader is directed to the several examples in the literature of its use in TB outbreak investigations [Gardy *et al.* 2011; Stucki *et al.* 2015; Lee *et al.* 2015; Torok *et al.* 2013; Kato-Maeda *et al.* 2013; Schurch *et al.* 2010; Ocheretina *et al.* 2015; Walker *et al.* 2013; Roetzer *et al.* 2013]. It is worth noting at this point that genotyping is occasionally required for clinical care, for instance, to rule out laboratory cross-contamination as a false-positive cause of a positive culture, or when trying to determine when a TB recurrence is due to relapse of the original infection *versus* exogenous reinfection. For both of these applications, the lessons of outbreak investigation indicate that WGS has higher resolution than traditional typing methods, such as spoligotyping, mycobacterial interspersed repetitive units (MIRUs), or restriction fragment length polymorphism (RFLP) [Gardy *et al.* 2011; Lee *et al.* 2015; Walker *et al.* 2013; Roetzer *et al.* 2013]. Therefore, it can be inferred that, for both situations, if the traditional method returns a result of 'different strain', WGS is not necessary to answer the clinical question. If, however, the traditional typing method returns a matched pattern, WGS may be required to confidently

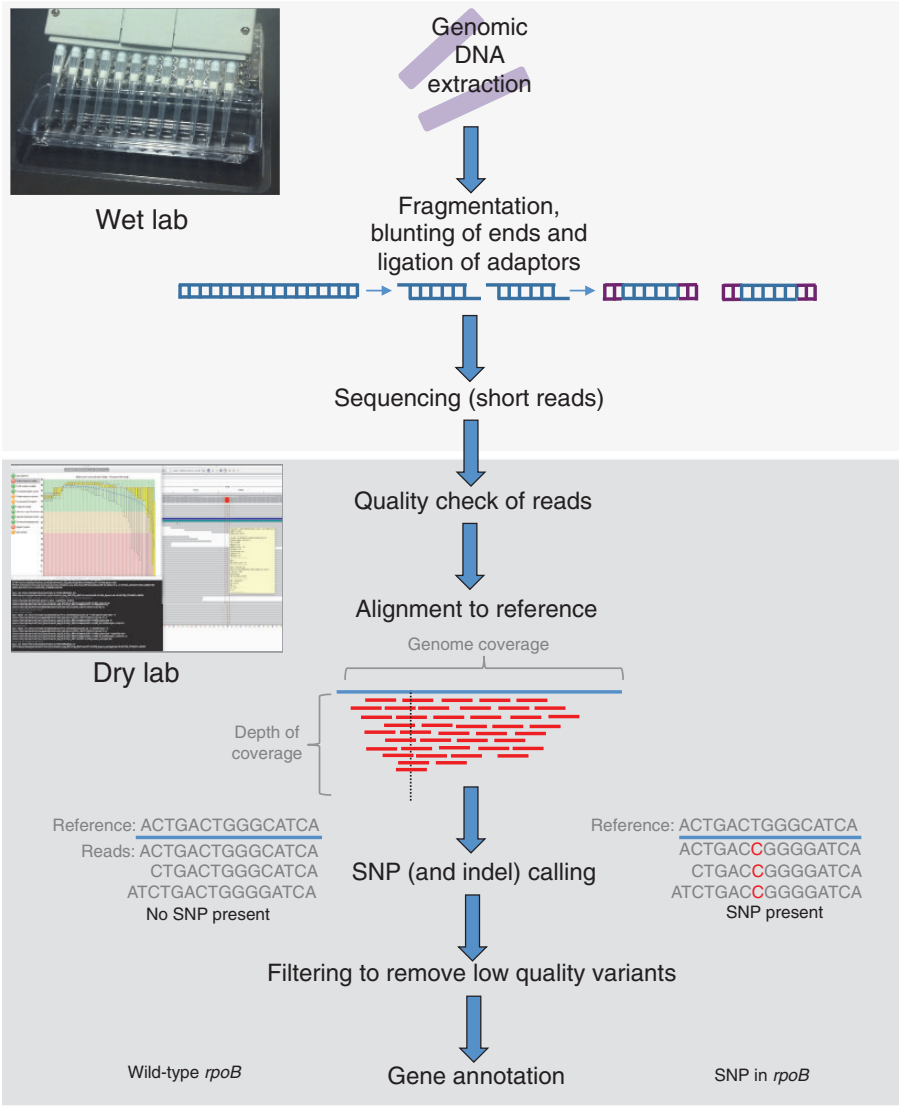


Figure 1. WGS workflow for *Mycobacterium tuberculosis*.
In brief, whole-genome sequencing (WGS) begins in the wet lab (top panel), wherein genomic DNA (gDNA) is extracted. For a *M. tuberculosis* culture, this is done in a biosafety level 3 laboratory. After DNA extraction, library preparation is conducted, wherein genomic DNA is fragmented into pieces. Uneven ends of gDNA are blunted and adaptor sequences are added. After passing quality control, libraries are advanced to sequencing. Further analysis occurs in the dry lab (bottom panel). Potential contamination is assessed and the quality of sequencing is evaluated on a per isolate basis, including the examination of Phred quality scores of the sequenced bases (where $\text{Phred} = -10*\log P_{error}$). FastQC, for example, is a software that can be used for such quality control, and is applied directly on raw sequence data (available from http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/, shown in the screenshot). Adaptors (and potentially low-quality base pairs) are trimmed and reads of length under a prespecified limit (e.g. 70 base pairs used by the 1000 Genomes Project) may be excluded (not shown). High-quality reads are aligned to a reference genome (this can be visualized in Integrative Genomics Viewer, also shown in screenshot [Thorvaldsdottir *et al.* 2013]), and metrics such as genome coverage (the percentage of the reference genome that has at least one read mapped to it) and depth of coverage (the average number of reads mapped to each locus) are evaluated. Isolates are retained if *a priori* quality measures are met. Reads are excluded if they map to more than one locus in the genome, and additional quality measures may be applied such as removing polymerase chain reaction duplicates and local realignment around indels. Once quality control steps are conducted, single-nucleotide polymorphisms and indels can then be 'called' compared with the reference genome. Low-quality variants are then removed using various filtering parameters to reduce the number of false positives. Genes are then annotated and repetitive regions and mobile elements may be filtered out of further analyses.

distinguish a related strain due to ancestry from a true match, with the latter being observed during laboratory cross-contamination or relapse.

Regardless of the application, the quality of WGS data depends on a number of factors, including the desired length of the sequencing reads and the cycle time [Quick *et al.* 2015]. These parameters in turn affect the turnaround time for results. Considering the most frequently used bench-top sequencers, raw sequencing results can be available in a clinically attractive span of just a few hours (for the Ion Torrent PGM) to as much as 39 hours with MiSeq for paired end 250 bp reads. By adjusting the sequencing protocol for MiSeq, it may be feasible to reduce this time frame without affecting key inferences, such as species and strain assignments [Quick *et al.* 2015]. An important consideration when making such adjustments is the 'depth of coverage'; the more reads that span a position in the reference genome, the more support there is for the base identified. The optimal depth of coverage to detect clinically relevant variants needs to be determined.

Another factor influencing the time to obtain these data is whether samples are batched or run independently. According to Quick and colleagues [Quick *et al.* 2015], the MiSeq can sequence up to ~100 isolates simultaneously. In our experience, the MiSeq 250-bp paired-end sequencing can generate a minimum of 10 million reads; if 20× coverage is desired, only ~57 isolates of *M. tuberculosis* can be run simultaneously [Lander and Waterman, 1988]. A batched approach such as this is typical in research labs and is clearly less expensive on a per-unit basis, as running a single isolate would cost the same as the whole collection of samples. Unfortunately, waiting until a queue of specimens has accumulated is not ideal for clinical labs, which need to process samples immediately on arrival and send reports 24 hours a day. A newer method, the Nanopore MinIon (Oxford Nanopore Technologies, Oxford, UK), offers much promise in addressing this problem. The MinIon runs a single sample at a time and was able to correctly speciate two *Salmonella enterica* isolates as well as place them in epidemiologic context within 2 h [Quick *et al.* 2015]. Earlier diagnosis and detection of SNPs connoting drug resistance could allow for more rapid initiation of treatment, compared with waiting for results from a batched analysis. However, the advantage of rapid results offered by the MinIon is currently offset by high error rates as reported by [Laver *et al.* 2015; Mikheyev and Tin, 2014; Quick *et al.* 2015]. While sequencing chemistry is improving and bioinformatics approaches are being developed to increase accuracy [Jain *et al.* 2015], further studies are needed to evaluate this method. As of yet, the MinIon has not been utilized for *M. tuberculosis*. It might be that these different platforms offer complementary opportunities for the clinical lab, for instance by using the Nanopore technology to rapidly speciate pathogenic organisms and the MiSeq for ongoing epidemiologic surveillance.

WGS for detection of *M. tuberculosis*, including the prediction of drug resistance

In the clinical mycobacteriology lab, the goal is to secure a diagnosis of active TB and to provide clinicians with guidance on which antibiotics they should or should not prescribe for their patients. These two goals have classically been achieved with phenotypic tests, some dating to the 19th century. This begs the obvious question of whether WGS can help modernize the TB lab, with the goal of offering faster and more accurate results.

The current clinical workflow for detection of *M. tuberculosis* in Canada is illustrated in Figure 2. Variations of this pathway may be seen in comparable high-resource countries. For more detailed reviews of *M. tuberculosis* laboratory diagnosis, the reader is referred to the literature [Parrish and Carroll, 2008, 2011; Drobniewski *et al.* 2013; Noor *et al.* 2015]. In brief, specimens from TB suspects are sent for smear microscopy to ascertain the presence of acid-fast bacilli. This test identifies the most infectious patients (i.e. with 'smear-positive' disease) [Behr *et al.* 1999]. Results of smear microscopy should be available within 24 h of receipt [Parrish and Carroll, 2011], however this method has low sensitivity [Steingart *et al.* 2006a, 2006b] and cannot distinguish *M. tuberculosis* from non-TB mycobacterium. Regardless of the results of microscopic examination, the same specimens are processed for culture, as detailed by Parrish and Carroll [Parrish and Carroll, 2011]. The culture is usually done using both solid and liquid media (typically mycobacterial growth indicator tubes [MGITs]), with growth usually observed in 1–3 weeks, depending on the mycobacterial inoculum

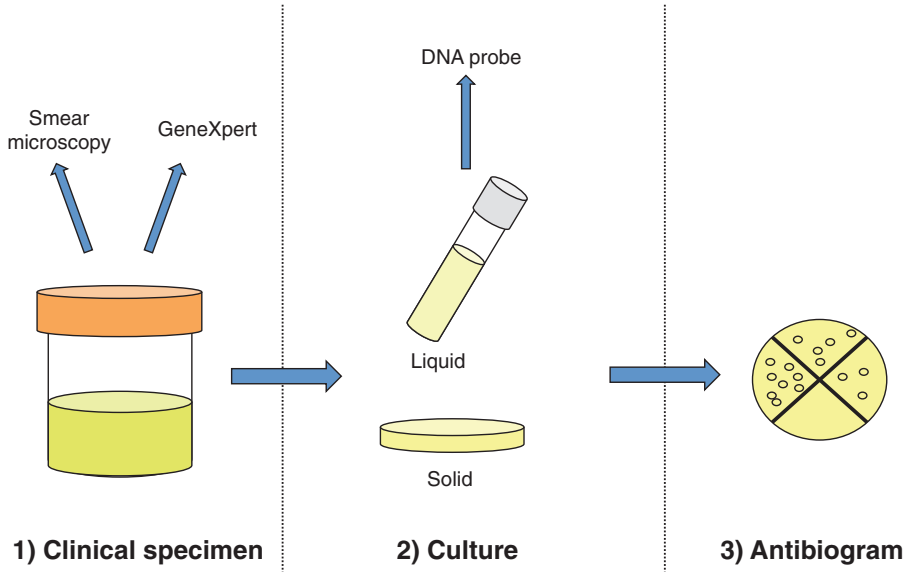


Figure 2. Clinical diagnostic workflow for *Mycobacterium tuberculosis*.
The three main steps in the current diagnostic workflow for *M. tuberculosis* are shown. As described in the text, whole-genome sequencing may have a potential role at each of these steps: (1) by being applied directly to the unprocessed clinical specimen or (2) by being conducted on the positive culture to predict drug resistance.

in the sample [Chihota *et al.* 2010; Fadzilah *et al.* 2009]. Once growth is observed (on solid media) or flagged by the machine (in the case of MGITs), a positive culture can be assigned a presumptive identification as *M. tuberculosis* complex using a DNA probe, usually within 24 h [Ichiyama *et al.* 1997]. Cultures are then sent to a reference laboratory for formal species confirmation and for drug susceptibility testing (DST) by phenotypic (i.e. growth-based) assays.

Superimposed on this classic workflow (smear microscopy, culture, then DST), laboratories have overlaid molecular testing over the past two decades, using a variety of different platforms and clinical strategies. The first molecular tests approved were only licensed for the speciation of smear microscopy-positive samples [Parrish and Carroll, 2011], so their key role was in assigning a microbial name to such a sputum sample [Vuorinen *et al.* 1995; Carpentier *et al.* 1995]. Then, with time and experience, it became recognized that nucleic acid amplification testing could be offered on smear-negative samples where there was a high clinical suspicion of TB [Centers for Disease Control and Prevention (CDC) 2009]. To reduce costs of controls, these 'rapid' first generation tests were generally batched and as a result, might only have been done twice or three times per week, depending on laboratory volume. More recently, the GeneXpert (Cepheid Inc., Sunnydale, CA, USA) has offered a random-access real-time nucleic acid amplification test, which can be done on a single sample, without having to wait for samples from other patients. GeneXpert is conducted directly on the clinical specimen to detect both the presence of *M. tuberculosis* DNA and mutations in the *rpoB* gene that predict resistance to the first-line drug, rifampin. In principle, results can be available in under 2 h [Boehme *et al.* 2010]. In practice, turn-around time depends on logistics; most testing is done in laboratories rather than clinics, necessitating delays due to shipping and handling [Alvarez *et al.* 2015]. The specificity of GeneXpert for *M. tuberculosis* detection is high, reported at >98%, but the sensitivity varies by smear status [Boehme *et al.* 2010; Steingart *et al.* 2014; Sohn *et al.* 2014], site (e.g. respiratory *versus* extrapulmonary) and type of sample (e.g. lymph node *versus* pleural [Maynard-Smith *et al.* 2014; Denkinger *et al.* 2014]). While GeneXpert is currently the fastest and arguably most useful diagnostic test in many parts of the world, it may be that its enduring legacy is catalyzing a paradigm shift away from phenotypic testing, towards genetic detection of *M. tuberculosis* as the primary

goal of the TB lab. If true, then the same pre-analytic principles (collecting sputum, delivering to lab, rendering the sample safe, extracting DNA) can serve as the basis for a more comprehensive interrogation of the mycobacterial genome, going beyond the *rpoB* gene to characterize the complete genome of the causative organism.

WGS for diagnosis

Until recently, the utility of WGS for *de novo* diagnosis of *M. tuberculosis* was unclear. WGS had relied exclusively on enriched DNA obtained from a pure bacterial culture, at which point the patient would have already been diagnosed. More recently, studies have examined the feasibility of sequencing *M. tuberculosis* directly from the clinical specimen [Doughty *et al.* 2014; Brown *et al.* 2015]. Sequencing eight smear positive samples, Doughty and colleagues obtained only 0.002× to 0.7× depth of coverage, with 20–99% of reads sequenced mapping to the human genome rather than *M. tuberculosis* [Doughty *et al.* 2014]. Brown and colleagues obtained similar results when sequencing directly from clinical samples, but when an oligonucleotide enrichment protocol was applied, they were able to obtain at least 20× depth of coverage on 20/24 smear positive, culture positive isolates, providing sufficient sequence depth to confidently speciate the organism present [Brown *et al.* 2015].

If WGS is to be applied on the patient sample, the conceptual advantage is a more rapid result. However, the vast majority of samples are negative for *M. tuberculosis*, even in a high-incidence setting [Demers *et al.* 2012], so some form of triage is needed to select the samples most likely to benefit from direct WGS. Furthermore, sputum is contaminated with host and other bacterial DNA, complicating bioinformatic analyses and reducing the overall depth of coverage obtained for the *M. tuberculosis* genome [Doughty *et al.* 2014]. While low coverage may not preclude the ability to confidently detect *M. tuberculosis*, it could seriously undermine the capacity to detect mutations associated with drug resistance (as shown by Doughty and colleagues [Doughty *et al.* 2014]), where the greatest clinical value of WGS may lie. In sum, these studies provide proof-of-principle that WGS of *M. tuberculosis* directly from clinical specimens is feasible, but the cost of the enrichment protocol (USD$350 per sample), the requirement for technical expertise and equipment, and the need for real-time bioinformatics to convert sequence files into clinically meaningful lab reports all present challenges to WGS supplanting smear microscopy and nucleic acid amplification as the primary test performed on clinical specimens.

If instead WGS is applied on the positive culture, then the benefit of rapidity has been lost, as the patient should already be isolated and started on treatment, based on either smear microscopy, a nucleic acid amplification test or the Accuprobe result on the culture. In this case, WGS may offer a different opportunity, which is a more rapid identification of antibiotic resistance.

WGS for resistance

In 2013, 3.5% of incident TB cases worldwide (95% confidence interval [CI] 2.2–4.7%) were estimated to have multidrug-resistant (MDR) TB, with an enrichment to 20.5% in cases with previous treatment (95% CI 13.6–27.5%) [World Health Organization, 2015]. As there is no evidence for ongoing acquisition of foreign DNA by *M. tuberculosis*, resistance occurs due to mutations in the chromosomal DNA, some of which have been mapped and mechanistically linked to the resistance phenotype [Nebenzahl-Guimaraes *et al.* 2014]. Phenotypic testing of a positive culture (called indirect DST) is the current gold standard for *M. tuberculosis*. The need for level 3 containment facilities and the requirement to perform an appropriate number of tests to maintain competence, however, have conspired to direct this most clinically meaningful assay to reference labs, entailing delays due to transport and handling. Therefore, while it is stated that first-line susceptibility results can be obtained in 2-4 weeks [Perkins and Cunningham, 2007; Migliori *et al.* 2008], such estimates reflect the time for work to be performed in the reference lab. When considering the time from sample acquisition to a final report, others provide longer timeframes, up to 2 months [Parrish and Carroll, 2008]. Until this information is available, the clinician faces an immediate dilemma, which is: 'What do I prescribe now?'. Inappropriate treatment risks generating further drug resistance, but delaying treatment until a final report is provided risks deleterious treatment outcomes [Park *et al.* 1996]. While one option is to attempt phenotypic testing directly on the patient sample (called 'direct DST'), there are still delays with the time to obtaining cultures, and susceptibility testing on the sputum sample brings its own challenges, since it is difficult to standardize the

Table 1. Examples of molecular diagnostics for drug resistance in *M. tuberculosis*.

Molecular test	Drug	Gene(s) targeted	Test performed on?	Sensitivity	Specificity	Turnaround time	Publication Type	Reference
GeneXpert (Cepheid Inc., Sunnydale, CA, USA)	Rifampin	*rpoB*	Directly on clinical specimen (sputum + other respiratory specimens)	95 (95% CrI 90–97); range 33–100	98 (95% CrI 97–99); range 83–100	1 h–4 days (run time <2 h)	Meta-analysis	[Steingart *et al.* 2014]
MTBDR*plus* (Hain Lifescience, Nehren, DE)	Rifampin	*rpoB*	Combined data for DNA from clinical specimen (respiratory + non-respiratory samples) + purified DNA from culture	98.4 (95% CI 95.1–99.5-); range 94–100*	98.9 (95% CI 96.8–99.7); range 95–100*	6 h–2 days*	Meta-analysis	[Ling *et al.* 2008]
	Isoniazid	*katG, inhA*	Combined data for DNA from clinical specimen (respiratory + non-respiratory samples) + purified DNA from culture	88.7 (95% CI 82.4–92.8-); range 57–100*	99.2 (95% CI 95.4–99.8); range 92–100*	6 h–2 days*	Meta-analysis	[Ling *et al.* 2008]
INNO-LiPA RifTB (Innogenetics), Zwijndrecht, Belgium	Rifampin	*rpoB*	DNA from clinical specimen (including non-respiratory)	Range 80–100%	All 100%	Not reported	Meta-analysis	[Morgan *et al.* 2005]
			Purified DNA from culture	Range 82–100%	Range 92–100%	Not reported	Meta-analysis	[Morgan *et al.* 2005]
MTBDR*sl* (Hain Lifescience, Nehren, DE)	Fluoroquinolones (including oxyfloxacin and levofloxacin)	*gyrA*	DNA from clinical specimen (smear positive sputum)	85.1 (95% CI 71.9–92.7-); range 50–100	98.2 (96.8–99.0); range 91–100	8 h–2 days, two studies	Meta-analysis	[Theron *et al.* 2014a]
			Purified DNA from culture	83.1 (95% CI 78.7–86.7-); range 57–100	97.7 (95% CI 94.3–99.1); range 77–100	1 day (after first line)–10 days, two studies	Meta-analysis	[Theron *et al.* 2014a]
	Aminoglycosides (including kanamycin, amikacin, capreomycin)	*rrs*	DNA from clinical specimen (smear positive sputum)	94.4 (95% CI 25.2–99.9-); range 9–100	98.2 (95% CI 88.9–99.7); range 67–100	8 h–2 days, two studies	Meta-analysis	[Theron, *et al.* 2014a]
			Purified DNA from culture	76.9 (95% CI 61.1–87.6-); range 25–100	99.5 (95% CI 97.1–99.9); range 86–100	1 day (after first line)–10 days, two studies	Meta-analysis	[Theron *et al.* 2014a]
	Ethambutol	*embB,*	DNA from clinical specimen (sputum)	55 (95% CI 47–63)	78 (95% CI 69–85)	Not reported	Meta-analysis	[Cheng *et al.* 2014]
			Purified DNA from culture	64 (95% CI 60–67)	70 (95% CI 67–74)	Not reported	Meta-analysis	(Cheng *et al.* 2014) (continued)

Table 1. Continued

Molecular test	Drug	Gene(s) targeted	Test performed on?	Sensitivity	Specificity	Turnaround time	Publication Type	Reference
AID TB Resistance (AID Diagnostika, Strassberg, DE)	Rifampin	*rpoB*	DNA from clinical specimen (respiratory, 95% smear positive)	100 (95% CI 89.8–99.0)	100 (95% CI 77.1–100)	Not reported, 'similar to MTBDR*plus*/MDRTB*sl*'	Individual study	[Molina-Moya *et al.* 2015]
			DNA from clinical specimen (respiratory + nonrespiratory, smear positive)	***	100 (95% CI 95.9–100)	<1 day	Individual study	[Ritter *et al.* 2014]
			MGIT culture	100% (95% CI 29–100)	100% (95% CI 92–100)	<1 day	Individual study	[Ritter *et al.* 2014]
	Isoniazid	*katG, inhA*	DNA from clinical specimen (respiratory, 95% smear positive)	97.8 (95% CI 87.0–99.9)	100 (95% CI 73.2–100)	Not reported, 'similar to MTBDR*plus*/MDRTB*sl*'	Individual study	[Molina-Moya *et al.* 2015]
			DNA from clinical specimen (respiratory + nonrespiratory, smear positive)	***	100 (95% CI 95.9–100)	<1 day	Individual study	[Ritter *et al.* 2014]
			MGIT culture	100% (95% CI 29–100)	100% (95% CI 92–100)	<1 day	Individual study	[Ritter *et al.* 2014]**
	Fluoroquinolones	*gyrA*	DNA from clinical specimen (respiratory, 95% smear positive)	33.3 (95% CI 6.0–75.9)	98.1 (95% CI 88.6–99.9)	Not reported, 'similar to MTBDR*plus*/MDRTB*sl*'	Individual study	[Molina-Moya *et al.* 2015]
			DNA from clinical specimen (respiratory + nonrespiratory, smear positive)	No resistance	100 (95% CI 88–100)	<1 day	Individual study	[Ritter *et al.* 2014]
	Ethambutol	*embB*	DNA from clinical specimen (respiratory, 95% smear positive)	60.0 (95% CI 42.2–75.6)	91.7 (95% CI 71.5–98.5)	Not reported, 'similar to MTBDR*plus*/MDRTB*sl*'	Individual study	[Molina-Moya *et al.* 2015]
			DNA from clinical specimen (respiratory + nonrespiratory, smear positive)	100 (95% CI 3–100)	100 (95% CI 87.7–100)	<1 day	Individual study	[Ritter *et al.* 2014]
	Aminoglycosides (kanamycin and capreomycin)	*rrs*	DNA from clinical specimen (respiratory, 95% smear positive)	100 (95% CI 77.1–100)	100 (95% CI 87.4–100)	Not reported, 'similar to MTBDR*plus*/MDRTB*sl*'	Individual study	[Molina-Moya *et al.* 2015]
			DNA from clinical specimen (respiratory + nonrespiratory, smear positive)	-	100 (95% CI 89.7–100)	<1 day	Individual study	[Ritter *et al.* 2014]
	Streptomycin	*RpsL, rrs*	DNA from clinical specimen (respiratory, 95% smear positive)	100 (95% CI 81.5–100)	96.6 (95% CI 80.4–99.8)	Not reported, 'similar to MTBDR*plus*/MDRTB*sl*'	Individual study	[Molina-Moya *et al.* 2015]

(continued)

Table 1. Continued

Molecular test	Drug	Gene(s) targeted	Test performed on?	Sensitivity	Specificity	Turnaround time	Publication Type	Reference
			DNA from clinical specimen (respiratory+nonrespiratory, smear positive)	–	100 (95% CI 89.7–100)	<1 day	Individual study	[Ritter *et al.* 2014]

For meta-analyses: if available, ranges are shown in addition to pooled estimates, to indicate potential heterogeneity. All tests shown, with exception of GeneXpert, are line probe assays. Where no sensitivity is reported, no isolates were identified with resistance to the target drug. *Includes studies using MTBDR (first-generation). **Results not shown for second-line drugs, as only testing was only conducted on the 3 samples with resistance to first-line drugs. AID predicted 3/3 isolates to be susceptible to second-line, confirmed with phenotypic DST. ***Study did not report separate test results for positive RIF and INH resistance.
CI, confidence interval; CrI, credible interval; INH, isoniazid; RIF, rifampin.

inoculum for such assays. It is at this moment of indecision that a molecular test could provide the most immediate clinical guidance, as exemplified by the GeneXpert test. For examples of molecular tests, along with sensitivity and specificity for respective drugs, see Table 1.

As most rifampin-resistant isolates are also isoniazid-resistant, the GeneXpert uses *rpoB* mutations associated with rifampin-resistance as a proxy for multi-drug resistance. However, not all rifampin-resistant organisms are isoniazid-resistant (i.e. there can be rifampin mono-resistance) and indeed, not all isolates predicted to be rifampin-resistant are confirmed on phenotype-based testing [Steingart *et al.* 2014]. In addition, not all rifampin-resistant isolates are detected based on the currently assessed mutations [Sanchez-Padilla *et al.* 2015; Jamieson *et al.* 2014a]. Finally, GeneXpert may fail to detect hetero-resistance, i.e. resistance-connoting mutations present in subpopulations within the patient [Zetola *et al.* 2014]. For all of these reasons, a broader-based assay, such as WGS, could offer the greatest clinical utility at this point in the diagnostic process, by looking beyond the targets of the current molecular assays.

By sequencing the whole genome, in theory all resistance-connoting mutations that can guide clinical treatment can be identified by comparing the genome of the patient isolate with detailed databases of known resistance markers [Sandgren *et al.* 2009; Flandrois *et al.* 2014]. In practice, this will work, if (a) these markers accurately predict *in vitro* phenotypic resistance, and (b) these markers predict clinical outcome. For the latter, we are unaware of studies that have directly assessed the utility of WGS data for predicting patient response to treatment. For the proximal goal of linking WGS to phenotypic resistance, there are emerging data which present a mixed message. Using online databases, supplemented with an updated search of the literature, Coll and colleagues [Coll *et al.* 2015] developed a mutation library and examined the concordance between genotypic predictions and phenotypic data for 788 isolates from diverse geographic settings. Among the drugs with sufficient phenotypic data (rifampin (RIF), isoniazid (INH), ethambutol (EMB), pyrazinamide (PZA) and streptomycin (STR)) as well as second-line drugs (amikacin (AMK), capreomycin (CAP), ethionamide (ETH), kanamycin (KAN), moxifloxicin (MOX), ofloxacin (OFX)), the sensitivity

of WGS for predicting resistance was highest for INH and RIF at 92.8% (95% CI 89.9–95.7) and 96.2 (95% CI 93.9–98.5). At the other end of the spectrum, the sensitivity of WGS for PZA-resistance was only 70.9% (95% CI 62.4–79.4). Thus, if WGS replaced phenotypic testing, one-twelfth of INH-resistant and one-third of PZA-resistant cases would receive these potentially hepatotoxic drugs, with little or no benefit. Specificity of WGS was highest for INH and RIF at 100% (95% CI 100–100%) and 98.1% (95% CI 96.8–99.4%), respectively, but for other drugs, specificity was as low as 81.7% (EMB).

In the same manuscript [Coll *et al.* 2015], Coll and colleagues also compared the performance of their database with KvarQ, a software that uses pre-specified 'testsuites' of known resistance-connoting mutations and other regions of interest to predict resistance [Steiner *et al.* 2014]. Using phenotypic data as the gold standard, sensitivity was substantially lower for nearly all drugs using the KvarQ method (though 95% CIs overlapped for all except EMB and KAN). Among first-line drugs, only RIF yielded similar point estimates to those obtained with Coll and colleagues' mutation library, with sensitivity of 95.8% (95% CI 93.4–98.2%), while sensitivity for INH was only 86.9% (95% CI 83.1–90.7%). No results were available for ETH and CAP using the KvarQ software. Specificity was generally higher using KvarQ, though this difference was only significant for EMB and STR. Specificity for RIF was similar to that obtained with the mutation database, at 97.9% (95% CI 96.5–99.3%).

In a similar study [Walker *et al.* 2015], Walker and colleagues selected 23 candidate resistance-associated genes from the literature [Sandgren *et al.* 2009] and then used an algorithm to characterize mutations (SNPs and indels) within these genes and their promoter regions as resistance-connoting or benign. In a training dataset of 2099 isolates, 120 resistance-connoting mutations were identified, 772 were classified as benign and 101 could not be classified as either (called 'uncharacterized'). The resistance-connoting and benign mutations identified in this training dataset were then used in a validation study on an additional 1552 genomes, 29% of which were resistant to at least one drug on drug susceptibility testing (DST). Using these mutations, authors were able to predict 89.2% of phenotypes as resistant or susceptible. 10.8% of phenotypes could not be predicted, as these contained mutations that had not been characterized. Among those where phenotype could be predicted and considering predictions for each drug independently, 112 of 6892 with drug-sensitive DST were predicted to be resistant based on WGS (1.6%), while 94 of 1221 with drug-resistant DST were erroneously predicted to be drug-sensitive (7.7%). The latter may be due to mutations with unknown function outside the 23 candidate genes interrogated. This is similar to Farhat and colleagues [Farhat *et al.* 2013]; in this study, authors performed targeted deep sequencing of known resistance genes to verify that resistance mutations were absent in subpopulations within isolates. They found that 13/47 isolates with phenotypic resistance had no previously known mutations. Unexplained resistance, wherein phenotypic resistance is present but known resistance-connoting mutations are absent has been most pronounced for PZA [Hewlett *et al.* 1995] and second-line drugs. For example, Farhat and colleagues [Farhat *et al.* 2013] found that, among isolates resistant to ciprofloxacin, KAN and CAP, 2/3, 6/18 and 1/6 isolates, respectively, had unexplained resistance. As the reliability of phenotypic testing is least well established for these drugs [Horne *et al.* 2013], this is where there is the greatest need for WGS, but presently also the greatest knowledge gap.

In clinical medicine, the physician wants to know whether the isolate has a resistance-connoting mutation or not, so that treatment can be tailored accordingly. Indeterminate test results offer little clinical guidance, and often steer clinicians to other antibiotics, where feasible. While it is logical to exclude isolates with uncharacterized mutations from a scientific paper that aims to understand resistance, in a clinical laboratory, these have to reported one way or the other. Analyses that classified such uncharacterized mutations as predictive of phenotypic susceptibility greatly affected test parameters; the sensitivity of WGS for INH and RIF resistance dropped from 94.2% (95% CI 91.1–96.5%) and 96.8% (95% CI 94.1–98.5%) with uncharacterized mutations excluded to 85.2% (95% CI 81.1–88.7%) and 91.7% (95% CI 87.9–94.5%) with uncharacterized mutations included, respectively. Sensitivity for PZA resistance in the latter analysis was the lowest overall, at only 24% (95% CI 17.9–30.9%). Until such mutations can be confidently assigned to the appropriate phenotype, it would seem that

parallel, or at the least, sequential phenotypic testing should remain part of the diagnostic pathway.

Furthermore, these publications generally included biased samples, with relatively high proportions of drug-resistant isolates. As many clinical labs identify primarily drug-sensitive isolates, the operating parameters of WGS for this purpose may change when evaluated against more representative samples. While authors had generally high specificity for most drugs, the predictive value depends on the underlying prevalence of drug resistance. In a country such as Canada, which detected RIF resistance among only 17 of 1380 *M. tuberculosis* complex isolates analyzed in 2013 [Public Health Agency of Canada, 2015], a specificity of 98.1–99.2% and sensitivity of 91.7–96.2% based on the results of Coll and colleagues [Coll *et al.* 2015] and Walker and coworkers [Walker *et al.* 2015] would equate to ~18 false positives per year, with a positive predictive value of only ~46%. Without subsequent phenotypic testing, these cases would be subject to second-line treatment, with prolonged, unnecessary hospitalization. Thus, WGS may be best reserved only for individuals in which there was a higher pretest probability of resistance (based on some *a priori* criteria for the use of WGS, e.g. previous treatment).

Despite these limitations, it is clear that WGS offers magnitudes more information than the molecular methods listed in Table 1, with the potential of greatly advancing clinical diagnostics for *M. tuberculosis*. While the WGS database of Coll and colleagues [Coll *et al.* 2015] performed similarly to GeneXpert for RIF resistance, it also allowed for determination of INH mutations, and had an overall accuracy of 95.8%, as compared to 93.1% for MTBDR*plus* (Hain Lifescience, Nehren, DE) ($p < 0.0004$). Accuracy was also higher for second-line drugs compared with MTBDR*sl* (Hain Lifescience, Nehren, DE) (96.3% *versus* 93.7%, $p < 0.0047$). Walker and colleagues [Walker *et al.* 2015] showed similar sensitivity and specificity of their algorithm for determining the correct phenotype using WGS as the collective results of MTBDR*plus*, MTBDR*sl* and AID (AID Diagnostika, Strassberg, DE) line probe assays (LPAs). In addition, while synonymous SNPs can present as false positives on both LPA or GeneXpert, Walker and colleagues were able to classify these as benign.

Overall, these data support the great potential of WGS as a tool to predict resistance. However, databases of *M. tuberculosis* genomes, along with associated phenotypic data, are essential to identify unrecognized and emerging mutations. In addition, our ability to accurately predict phenotypic resistance is limited by our understanding of epistasis (the interaction between mutations, which can influence phenotype [Trauner *et al.* 2014]); mutations associated with resistance have been found in phenotypically sensitive bacteria [Walker *et al.* 2015], in some cases potentially due to interaction with other mutations in the genome. Until additional data are gathered, it can be foreseen that WGS may serve as an added, rather than a replacement test, on the diagnostic pipeline (Figure 2). This would incur added costs to the lab, something that is clearly less attractive than WGS simply replacing drug susceptibility testing (DST), with all its labor and reagent costs. One need look no further than the example of HIV treatment to imagine a world where genotype-based data are used to predict drug resistance, and hence treatment decisions. However, for all of the aforementioned reasons, we submit that reference labs need to maintain competence in phenotypic DST for the foreseeable future.

Another issue for clinical application of WGS is timeliness of reporting. As of yet, two papers reported on the application of WGS in 'real-time' to clinical cases: a case report of a patient [Köser *et al.* 2012] with extremely drug-resistant (XDR) TB (defined as MDR TB plus resistance to an injectable second-line drug and a fluoroquinolone) and a prospective cohort of patients in the United Kingdom suspected of having XDR TB [Witney *et al.* 2015]. Köser and colleagues successfully obtained sequence data from a 3-day-old MGIT culture, identifying two concurrent but distinct strains of *M. tuberculosis* [Köser *et al.* 2012]. Predicted resistance and sensitivity concurred with phenotypic results for all drugs tested, while WGS predicted resistance to an additional five drugs. While WGS results had no impact on treatment, WGS did identify a mutation in the gene activating PAS in the minority strain, despite a phenotypic determination of PAS-S. Unfortunately, the functional impact of this was unknown. Witney and colleagues [Witney *et al.* 2015] selectively applied WGS to six cases with potential XDR TB, identified over 6 years in London, with multiple isolates sequenced per patient. Results for five out of six

cases were available in a clinically actionable time frame. Genotypic and phenotypic resistance were 100% concordant for INH and RIH, while discrepancies were reported in PZA, EMB, fluoroquinolones (OFX and MOX), AMK, KAN, CAP, PRO and PAS. In terms of clinical utility, WGS data helped guide treatment decisions by confirming PZA resistance in one case, and refuting an XDR diagnosis in favor of MDR in another. For another case, clinicians decided to continue with treatment with EMB, despite development of phenotypic resistance, as WGS failed to identify mutations in *embA* or *embB* that could explain the change in DST.

The Witney and colleagues study also illustrated that for WGS data to be used clinically, the results need to be analyzed rapidly and presented in a clear, easily interpretable manner. Several groups have produced online tools (e.g. 'PhyResSE' [Feuerriegel *et al.* 2015] and 'TB Profiler' [Coll *et al.* 2015]) wherein raw sequencing data for an isolate can be uploaded and analyzed for resistance-connoting mutations. As mentioned previously, the KvarQ software can also predict resistance from raw sequencing data; in contrast to PhyResSE and TB Profiler, this can be done on a local server [Steiner *et al.* 2014]. Yet, despite efforts to make these reports accessible to the wider scientific community, a knowledge of genomics and/or bioinformatics is still required to interpret results. As an example, the quality of SNPs is provided with details such as depth of coverage, a parameter that most clinicians would be uncomfortable judging. Presently, PhyResSE and TB profiler are explicitly for research purposes only, which poses regulatory hurdles to the delivery of results destined for the clinical chart. Witney and colleagues [Witney *et al.* 2015] piloted a WGS report during the course of their study, but, unfortunately, clinician perception of this report and its interpretability was not assessed. Furthermore, though 'best practices' have been proposed for identifying SNPs [Olson *et al*. 2015], the current bioinformatics workflows used to analyze WGS data remain largely unstandardized. For implementation in the clinical lab, appropriate quality control measures [Clinical and Laboratory Standards Institute, 2014] and a standardized workflow need to be established. The lessons of the past five decades of emerging antibiotic resistance have demonstrated that even a simple dichotomous test result, i.e. resistant or susceptible, does not always predict appropriate care. Therefore, the application of WGS-based results to clinical care may benefit from evaluations done by experts in implementation science, rather than genomics or microbiology.

Conclusion

Offering increased resolution and substantially more data compared with conventional methods, WGS has revolutionized the arena of molecular epidemiology. Now, it seems poised to do the same for the clinical microbiology laboratory. The appeal of WGS for *M. tuberculosis* (and other pathogens) lies in the quantity of data provided; with one test, an organism can be speciated, resistance mutations can be detected and the strain can be placed in the context of the local epidemiology. The challenge of WGS also lies in the quantity of data provided; the same test can occupy a team of bioinformaticians, yet generate results that few clinicians can currently interpret. Furthermore, for WGS data to be clinically useful, results must be available in sufficient time to guide patient care. Recent advances such as sequencing directly from clinical samples and the rapid workflow of the Nanopore MinIon may facilitate this. The decision to whom this 'test' will be applied is also critical. Though no studies to date have examined cost-effectiveness of implementing WGS, it can be predicted that application of this test to all, unselected samples without removing other steps in laboratory workflow could be prohibitively expensive. Therefore, it can be foreseen that WGS will be applied selectively, for instance, on patients with Rifampin resistance mutations detected by the GeneXpert assay.

The issues raised above are only further amplified when contemplating the countries of the world that suffer the greatest burden of TB and have the highest prevalence of drug-resistant strains. While it is clearly feasible to ship sequencing machines around the world, as has already been done with the GeneXpert platform, it is not as simple to distribute the technical and bioinformatic expertise required for next-generation sequencing where it is needed. A potential solution to the latter is open-source coding and online data treatment, but this is currently lacking for clinical use, even in settings with expertise in these methods. Ultimately, what is needed is an easy-to-use software complete with a graphical user interface that is capable of converting data-intense sequence files into a simple, concise clinical message. As done with GeneXpert [Theron *et al.* 2014b], these outputs then need to be field-

tested in settings with a sufficient burden of drug-resistant TB to enable evaluation of whether test results altered treatment decisions and clinical outcomes. The relatively small number of MDR TB patients in countries such as Canada may preclude a formal evaluation of patient outcomes, simply due to sample size considerations. In order to assess its clinical utility for resource-rich countries where its use has been pioneered, we may need to first embed WGS in treatment studies conducted in the developing world, where the challenge posed by TB and drug resistance remains the greatest.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: *M. tuberculosis* genomic epidemiology work in the laboratory of MAB is supported by the Canadian Institutes of Health Research (MOP# 125858).

Conflict of interest statement

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Altschul, S., Gish, W., Miller, W., Myers, E. and Lipman, D. (1990) Basic local alignment search tool. *J Mol Biol* 215: 403–410.

Alvarez, G., Van Dyk, D., Desjardins, M., Yasseen, A. III, Aaron, S., Cameron, D. *et al.* (2015) The feasibility, accuracy, and impact of Xpert MTB/RIF testing in a remote aboriginal community in Canada. *Chest* 148: 767–773.

Behr, M., Warren, S., Salamon, H., Hopewell, P., Ponce de Leon, A., Daley, C. *et al.* (1999) Transmission of *Mycobacterium tuberculosis* from patients smear-negative for acid-fast bacilli. *Lancet* 353: 444–449.

Boehme, C., Nabeta, P., Hillemann, D., Nicol, M., Shenai, S., Krapp, F. *et al.* (2010) Rapid molecular detection of tuberculosis and rifampin resistance. *N Engl J Med* 363: 1005–1015.

Brown, A., Bryant, J., Einer-Jensen, K., Holdstock, J., Houniet, D., Chan, J. *et al.* (2015) Rapid whole-genome sequencing of *Mycobacterium tuberculosis* isolates directly from clinical samples. *J Clin Microbiol* 53: 2230–2237.

Bryant, J., Schurch, A., van Deutekom, H., Harris, S., de Beer, J., de Jager, V. *et al.* (2013) Inferring patient to patient transmission of *Mycobacterium tuberculosis* from whole genome sequencing data. *BMC Infect Dis* 13: 110.

Carpentier, E., Drouillard, B., Dailloux, M., Moinard, D., Vallee, E., Dutilh, B. *et al.* (1995) Diagnosis of tuberculosis by Amplicor Mycobacterium-Tuberculosis Test - a multicenter study. *J Clin Microbiol* 33: 3106–3110.

Casali, N., Nikolayevskyy, V., Balabanova, Y., Harris, S., Ignatyeva, O., Kontsevaya, I. *et al.* (2014) Evolution and transmission of drug-resistant tuberculosis in a Russian population. *Nat Genet* 46: 279–286.

Centers for Disease Control and Prevention (CDC). (2009) Updated guidelines for the use of Nucleic Acid Amplification Tests in the diagnosis of tuberculosis. *MMWR Morb Mortal Wkly Rep* 58: 7–10.

Cheng, S., Cui, Z., Li, Y. and Hu, Z. (2014) Diagnostic accuracy of a molecular drug susceptibility testing method for the antituberculosis drug ethambutol: a systematic review and meta-analysis. *J Clin Microbiol* 52: 2913–2924.

Chihota, V., Grant, A., Fielding, K., Ndibongo, B., van Zyl, A., Muirhead, D. *et al.* (2010) Liquid *versus* solid culture for tuberculosis: performance and cost in a resource-constrained setting. *Int J Tuberc Lung Dis* 14: 1024–1031.

Clinical and Laboratory Standards Institute. (2014) Nucleic acid sequencing methods in diagnostic laboratory medicine; Approved guideline - second edition. CLSI document MM09-A2. Clinical and Laboratory Standards Institute: Wayne, PA.

Cole, S., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D. *et al.* (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393: 537–544.

Coll, F., McNerney, R., Preston, M., Afonso Guerra-Assuncao, J., Warry, A., Hill-Cawthorne, G. *et al.* (2015) Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. *Genome Med* 7: 51.

Comas, I., Chakravartti, J., Small, P., Galagan, J., Niemann, S., Kremer, K. *et al.* (2010) Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. *Nat Genet* 42: 498–503.

Demers, A., Verver, S., Boulle, A., Warren, R., van Helden, P., Behr, M. *et al.* (2012) High yield of culture-based diagnosis in a TB-endemic setting. *BMC Infect Dis* 12: 218.

Denkinger, C., Schumacher, S., Boehme, C., Dendukuri, N., Pai, M. and Steingart, K. (2014) Xpert MTB/RIF Assay for the diagnosis of extrapulmonary tuberculosis: a systematic review and meta-analysis. *Eur Respir J* 44: 435–446.

Doughty, E., Sergeant, M., Adetifa, I., Antonio, M. and Pallen, M. (2014) Culture-independent detection and characterisation of *Mycobacterium tuberculosis* and *M. Africanum* in sputum samples using shotgun metagenomics on a benchtop sequencer. *Peer J* 2(3): e585, DOI: 10.7717/peerj.585/supp-3.

Drobniewski, F., Nikolayevskyy, V., Maxeiner, H., Balabanova, Y., Casali, N., Kontsevaya, I. *et al.* (2013) Rapid diagnostics of tuberculosis and drug resistance in the industrialized world: clinical and public health benefits and barriers to implementation. *BMC Med* 11: 190.

Fadzilah, M., Peng Ng, K. and Fong Ngeow, Y. (2009) The manual MGIT system for the detection of *M. tuberculosis* in respiratory specimens: an experience in the University Malaya Medical Centre. *Malay J Pathol* 31: 93–97.

Farhat, M., Shapiro, B., Kieser, K., Sultana, R., Jacobson, K., Victor, T. *et al.* (2013) Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. *Nat Genet* 45: 1183–1189.

Feuerriegel, S., Schleusener, V., Beckert, P., Kohl, T., Miotto, P., Cirillo, D. *et al.* (2015) PhyResSE: a web tool delineating *Mycobacterium tuberculosis* antibiotic resistance and lineage from whole-genome sequencing data. *J Clin Microbiol* 53: 1908–1914.

Flandrois, J., Lina, G. and Dumitrescu, O. (2014) MUBII-TB-DB: a database of mutations associated with antibiotic resistance in *Mycobacterium tuberculosis*. *BMC Bioinform* 15: 107.

Gardy, J., Johnston, J., Ho Sui, S., Cook, V., Shah, L., Brodkin, E. *et al.* (2011) Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N Engl J Med* 364: 730–739.

Guerra-Assuncao, J., Crampin, A. and Houben, R. (2015) Large-scale whole genome sequencing of *M. tuberculosis* provides insights into transmission in a high prevalence area. *Elife* 4: e05166.

Harris, S., Torok, M., Cartwright, E., Quail, M., Peacock, S. and Parkhill, J. (2013) Read and assembly metrics inconsequential for clinical utility of whole-genome sequencing in mapping outbreaks. *Nat Biotechnol* 31: 592–594.

Heather, J. and Chan, B. (2015) The sequence of sequencers: The history of sequencing DNA. *Genomics* DOI: 10.1016/j.ygeno.2015.11.003.

Hewlett, D., Horn, D. and Alfalla, C. (1995) Drug-resistant tuberculosis: Inconsistent results of pyrazinamide susceptibility testing. *JAMA* 273: 916–917.

Horne, D., Pinto, L., Arentz, M., Lin, S., Desmond, E., Flores, L. *et al.* (2013) Diagnostic accuracy and reproducibility of WHO-endorsed phenotypic drug susceptibility testing methods for first-line and second-line antituberculosis drugs. *J Clin Microbiol* 51: 393–401.

Ichiyama, S., Iinuma, Y., Yamori, S., Hasegawa, Y., Simokata, K. and Nakashima, N. (1997) Mycobacterium growth indicator tube testing in conjunction with the AccuProbe or the AMPLICOR-PCR assay for detecting and identifying mycobacteria from sputum samples. *J Clin Microbiol* 35: 2022–2025.

Jain, M., Fiddes, I., Miga, K., Olsen, H., Paten, B. and Akeson, M. (2015) Improved data analysis for the MinION Nanopore sequencer. *Nat Meth* 12: 351–356.

Jamieson, F., Guthrie, J., Neemuchwala, A., Lastovetska, O., Melano, R. and Mehaffy, C. (2014a) Profiling of rpoB mutations and MICs for rifampin and rifabutin in *Mycobacterium tuberculosis*. *J Clin Microbiol* 52: 2157–2162.

Jamieson, F., Teatero, S., Guthrie, J.L., Neemuchwala, A., Fittipaldi, N. and Mehaffy, C. (2014b) Whole-genome sequencing of the *Mycobacterium tuberculosis* Manila sublineage results in less clustering and better resolution than Mycobacterial Interspersed Repetitive-Unit-Variable-Number Tandem-Repeat (MIRU-VNTR) typing and spoligotyping. *J Clin Microbiol* 52: 3795–3798.

Kato-Maeda, M., Ho, C., Passarelli, B., Banaei, N., Grinsdale, J., Flores, L. *et al.* (2013) Use of whole genome sequencing to determine the microevolution of *Mycobacterium tuberculosis* during an outbreak. *PLoS One* 8(3): e58235.

Köser, C., Holden, M., Ellington, M., Cartwright, E., Brown, N., Ogilvy-Stuart, A. *et al.* (2012) Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *N Engl J Med* 366: 2267–2275.

Kwong, J., McCallum, N., Sintchenko, V. and Howden, B. (2015) Whole genome sequencing in clinical and public health microbiology. *Pathology* 47: 199–210.

Lander, E. and Waterman, M. (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* 2: 231–239.

Laver, T., Harrison, J., O'Neill, P., Moore, K., Farbos, A., Paszkiewicz, K. *et al.* (2015) Assessing the performance of the Oxford Nanopore technologies MinIon. *Biomol Detect Quantif* 3: 1–8.

Lee, R., Radomski, N., Proulx, J., Manry, J., McIntosh, F., Desjardins, F. *et al.* (2015) Reemergence and amplification of tuberculosis in the Canadian arctic. *J Infect Dis* 211: 1905–1914.

Ling, D., Zwerling, A. and Pai, M. (2008) GenoType MTBDR assays for the diagnosis of multidrug-resistant tuberculosis: a meta-analysis. *Eur Respir J* 32: 1165–1174.

Loman, N., Constantinidou, C., Chan, J., Halachev, M., Sergeant, M., Penn, C. *et al.* (2012a) High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nat Rev Microbiol* 10: 599–606.

Loman, N., Misra, R., Dallman, T., Constantinidou, C., Gharbia, S., Wain, J. *et al.* (2012b) Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol* 30: 434–439.

Long, S., Beres, S., Olsen, R. and Musser, J. (2014) Absence of patient-to-patient intrahospital

transmission of *Staphylococcus aureus* as determined by whole-genome sequencing. *mBio* 5(5): e01692–e1714.

Maynard-Smith, L., Larke, N., Peters, J. and Lawn, S. (2014) Diagnostic accuracy of the Xpert MTB/RIF assay for extrapulmonary and pulmonary tuberculosis when testing non-respiratory samples: a systematic review. *BMC Infect Dis* 14: 709.

Migliori, G., Matteelli, A., Cirillo, D. and Pai, M. (2008) Diagnosis of multidrug-resistant tuberculosis and extensively drug-resistant tuberculosis: current standards and challenges. *Can J Infect Dis Med Microbiol* 19: 169–172.

Mikheyev, A. and Tin, M. (2014) A first look at the Oxford Nanopore MinION Sequencer. *Mol Ecol Resour* 14: 1097–1102.

Molina-Moya, B., Lacoma, A., Prat, C., Diaz, J., Dudnyk, A., Haba, L. *et al.* (2015) AID TB Resistance line probe assay for rapid detection of resistant *Mycobacterium tuberculosis* in clinical samples. *J Infect* 70: 400–408.

Morgan, M., Kalantri, S., Flores, L. and Pai, M. (2005) A commercial line probe assay for the rapid detection of rifampicin resistance in *Mycobacterium tuberculosis*: a systematic review and meta-analysis. *BMC Infect Dis* 5(1): 62.

Nebenzahl-Guimaraes, H., Jacobson, K., Farhat, M. and Murray, M. (2014) Systematic review of allelic exchange experiments aimed at identifying mutations that confer drug resistance in *Mycobacterium tuberculosis*. *J Antimicrobial Chemother* 69: 331–342.

Noor, K., Shephard, L. and Bastian, I. (2015) Molecular diagnostics for tuberculosis. *Pathology* 47: 250–256.

Ocheretina, O., Shen, L., Escuyer, V., Mabou, M., Royal-Mardi, G., Collins, S. *et al.* (2015) Whole genome sequencing investigation of a tuberculosis outbreak in Port-Au-Prince, Haiti caused by a strain with a 'Low-Level' rpoB mutation L511P – insights into a mechanism of resistance escalation. *PLoS One* 10(6): e0129207.

Olson, N., Lund, S., Colman, R., Foster, J., Sahl, J., Schupp, J. *et al.* (2015) Best practices for evaluating single nucleotide variant calling methods for microbial genomics. *Front Genet* 6: 235.

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M. *et al.* (2014) A survey of tools for variant analysis of next-generation genome sequencing data. *Brief Bioinform* 15: 256–278.

Park, M., Davis, A., Schluger, N., Cohen, H. and Rom, W. (1996) Outcome of MDR-TB patients, 1983–1993 - prolonged survival with appropriate therapy. *Am J Respir Crit Care Med* 153: 317–324.

Parrish, N. and Carroll, K. (2008) Importance of improved TB diagnostics in addressing the extensively drug-resistant TB crisis. *Future Microbiol* 3: 405–413.

Parrish, N. and Carroll, K. (2011) Role of the clinical mycobacteriology laboratory in diagnosis and management of tuberculosis in low-prevalence settings. *J Clin Microbiol* 49: 772–776.

Perkins, M. and Cunningham, J. (2007) Facing the crisis: improving the diagnosis of tuberculosis in the HIV era. *J Infect Dis* 196(Suppl. 1): S15–S27.

Price, J., Golubchik, T., Cole, K., Wilson, D., Crook, D., Thwaites, G. *et al.* (2014) Whole-genome sequencing shows that patient-to-patient transmission rarely accounts for acquisition of *Staphylococcus aureus* in an intensive care unit. *Clin Infect Dis* 58: 609–618.

Public Health Agency of Canada. (2015) Tuberculosis: Drug resistance in Canada. 2013, Minister of Public Works and Government Services Canada: Ottawa (Canada).

Quick, J., Cumley, N., Wearn, C., Niebel, M., Constantinidou, C., Thomas, C. *et al.* (2014) Seeking the source of *Pseudomonas aeruginosa* infections in a recently opened hospital: an observational study using whole-genome sequencing. *BMJ Open* 4: e006278.

Quick, J., Ashton, P., Calus, S., Chatt, C., Gossain, S., Hawker, J. *et al.* (2015) Rapid draft sequencing and real-time Nanopore sequencing in a hospital outbreak of Salmonella. *Genome Biol* 16: 114.

Ritter, C., Lucke, K., Sirgel, F., Warren, R., van Helden, P., Bottger, E. *et al.* (2014) Evaluation of the AID TB Resistance line probe assay for rapid detection of genetic alterations associated with drug resistance in *Mycobacterium tuberculosis* strains. *J Clin Microbiol* 52: 940–946.

Roetzer, A., Diel, R., Kohl, T., Rückert, C., Nübel, U., Blom, J. *et al.* (2013) Whole genome sequencing *versus* traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiological study. *PLoS Med* 10(2): e1001387.

Sanchez-Padilla, E., Merker, M., Beckert, P., Jochims, F., Diamini, T., Kahn, P. *et al.* (2015) Detection of drug-resistant tuberculosis by Xpert MTB/RIF in Swaziland. *N Engl J Med* 372: 1181–1182.

Sandgren, A., Strong, M., Muthukrishnan, P., Weiner, B., Church, G. and Murray, M. (2009) Tuberculosis drug resistance mutation database. *PLoS Med* 6(2): e1000002.

Schurch, A., Kremer, K., Daviena, O., Kiers, A., Boeree, M., Siezen, R. *et al.* (2010) High-resolution typing by integration of genome sequencing data in a large tuberculosis cluster. *J Clin Microbiol* 48: 3403–3406.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O. and Huttenhower, C. (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Meth* 9: 811–814.

SenGupta, D., Cummings, L., Hoogestraat, D., Butler-Wu, S., Shendure, J., Cookson, B. *et al.* (2014) Whole-genome sequencing for high-resolution

investigation of methicillin-resistant *Staphylococcus aureus* epidemiology and genome plasticity. *J Clin Microbiol* 52: 2787–2796.

Snitkin, E., Zelazny, A., Thomas, P., Stock, F., NISC Comparative Sequencing Program Group, Henderson, D. *et al.* (2012) Tracking a hospital outbreak of carbapenem-resistant *Klebsiella pneumoniae* with whole-genome sequencing. *Sci Transl Med* 4(148): 148ra116.

Sohn, H., Aero, A., Menzies, D., Behr, M., Schwartzman, K., Alvarez, G. *et al.* (2014) Xpert MTB/RIF testing in a low tuberculosis incidence, high-resource setting: limitations in accuracy and clinical impact. *Clin Infect Dis* 58: 970–976.

Steiner, A., Stucki, D., Coscolla, M., Borell, S. and Gagneux, S. (2014) KvarQ: targeted and direct variant calling from fastq reads of bacterial genomes. *BMC Genom* 15: 881.

Steingart, K., Schiller, I., Horne, D., Pai, M., Boehme, C. and Dendukuri, N. (2014) Xpert® MTB/RIF assay for pulmonary tuberculosis and rifampicin resistance in adults (review). *Cochrane Libr* 1: 1–168.

Steingart, K., Henry, M., Ng, V., Hopewell, P., Ramsay, A., Cunningham, J., Urbanczik, R. *et al.* (2006a) Fluorescence *versus* conventional sputum smear microscopy for tuberculosis: a systematic review. *Lancet Infect Dis* 6: 570–581.

Steingart, K., Ng, V., Henry, M., Hopewell, P., Ramsay, A., Cunningham, J., Urbanczik, R. *et al.* (2006b) Sputum processing methods to improve the sensitivity of smear microscopy for tuberculosis: a systematic review. *Lancet Infect Dis* 6: 664–674.

Stucki, D., Ballif, M., Bodmer, T., Coscolla, M., Maurer, A., Droz, S. *et al.* (2015) Tracking a tuberculosis outbreak over 21 years: strain-specific single-nucleotide polymorphism typing combined with targeted whole-genome sequencing. *J Infect Dis* 211: 1306–1316.

Theron, G., Peter, J., Richardson, M., Barnard, M., Donegan, S., Warren, R. *et al.* (2014a) The diagnostic accuracy of the GenoType MTBDR*sl* assay for the detection of resistance to second-line anti-tuberculosis drugs (review). *Cochrane Libr* 10: 1–123.

Theron, G., Zijenah, L., Chanda, D., Clowes, P., Rachow, A., Lesosky, M. *et al.* (2014b) Feasibility, accuracy, and clinical effect of point-of-care XpertMTB/RIF testing for tuberculosis in primary care settings in Africa: a multicentre, randomised, controlled trial. *Lancet* 383: 424–435.

Thorvaldsdottir, H., Robinson, J. and Mesirov, J. (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 14: 178–192.

Torok, M., Reuter, S., Bryant, J., Koser, C., Stinchcombe, S., Nazareth, B. *et al.* (2013) Rapid whole-genome sequencing for investigation of a suspected tuberculosis outbreak. *J Clin Microbiol* 51: 611–614.

Trauner, A., Borrell, S., Reither, K. and Gagneux, S. (2014) Evolution of drug resistance in tuberculosis: recent progress and implications for diagnosis and therapy. *Drugs* 74: 1063–1072.

Vuorinen, P., Miettinen, A., Vuento, R. and Hallstrom, O. (1995) Direct detection of *Mycobacterium tuberculosis* complex in respiratory specimens by Gen-Probe Amplified *Mycobacterium tuberculosis* Direct Test and Roche Amplicor *Mycobacterium Tuberculosis* Test. *J Clin Microbiol* 33: 1856–1859.

Walker, T., Ip, C., Harrell, R., Evans, J., Kapatai, G., Dedicoat, M. *et al.* (2013) Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *Lancet Infect Dis* 13: 137–146.

Walker, T., Kohl, T., Omar, S., Hedge, J., Del Ojo Elias, C., Bradley, P. *et al.* (2015) Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: a retrospective cohort study. *Lancet Infect Dis* 15: 1193–1202.

Witney, A., Gould, K., Arnold, A., Coleman, D., Delgado, R., Dhillon, J. *et al.* (2015) Clinical application of whole-genome sequencing to inform treatment for multidrug-resistant tuberculosis cases. *J Clin Microbiol* 53: 1473–1483.

Wood, D. and Salzberg, S. (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 15(3): R46.

World Health Organization. (2015) Global Tuberculosis Report 2014, World Health Organization: Geneva.

Zetola, N., Shin, S., Tumedi, K., Moeti, K., Ncube, R., Nicol, M. *et al.* (2014) Mixed *Mycobacterium tuberculosis* complex infections and false-negative results for rifampin resistance by GeneXpert MTB/RIF are associated with poor clinical outcomes. *J Clin Microbiol* 52: 2422–2429.