Investigating within-host diversity of *Mycobacterium tuberculosis* reveals novel super-spreaders in the Canadian North

Robyn S. Lee PhD[1,2]*, Jean-François Proulx MD[3], Fiona McIntosh BSc[4], Marcel A. Behr MD[4], William P. Hanage PhD[1,2]

1. Center for Communicable Disease Dynamics, Harvard T. H. Chan School of Public Health, 677 Huntington Avenue, Kresge Building, Room 506I, Boston, MA, USA, 02115
2. Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA, USA, 677 Huntington Avenue, Kresge Building, 5th floor, Boston, MA, USA, 02115
3. Nunavik Regional Board of Health and Social Services, Kuujjuaq, Québec, Canada
4. The Research Institute of McGill University Health Centre, Montréal, Québec, Canada

*Address correspondence to:
Dr. Robyn Lee, PhD
Harvard T.H. Chan School of Public Health
677 Huntington Avenue
Kresge Building, Room 506I
Boston, MA 02115
Email: robyn.s.c.lee@gmail.com
Phone: 617-432-7955

Word count: 3485
Figures: 1
Tables: 1

Research in Context.

Evidence before this study.

Modeling studies suggest that within-host diversity can influence our inferences of transmission. As whole genome sequencing is being implemented for routine surveillance of *M. tuberculosis* by major public health departments, including Public Health England and the Centers for Disease Control in the United States, it is imperative we determine how to detect such diversity and when how it can be used to inform our understanding of transmission. To identify relevant studies, we searched PubMed in June 26, 2019 using the terms ("within-host diversity" OR "microevolution") AND "sequencing" AND "transmission" AND "tuberculosis". No language restrictions were applied. Most studies identified (summarized in a recent systematic review) relied on sequencing repeat samples to detect diversity within a single host. Unfortunately, outside recurrence or treatment failure, repeat samples (either at the same or different timepoints) are seldom available for public health. Other studies, using single samples, investigated diversity by sequencing all colonies on a plate using a 'clean sweep' and looking for mixed alleles when calling variants. However, these studies typically sequenced to a depth of only ~50x depth; while standard for public health, this depth may not be sufficient to detect all epidemiologically-

important variation. The additional value of 'deep' sequencing has not been systematically examined.

Added value of this study. We investigate the application of deep sequencing (i.e., to ~500-1000x) for the investigation of within-host diversity. Using a previously well-described outbreak of TB, we illustrate that deep sequencing data can help identify novel super-spreaders that are not detectable using routine sequencing approaches (i.e., with ~50x depth), and more accurately resolve *Mycobacterium tuberculosis* transmission networks for public health. We also highlight a role for deep sequencing in discriminating relapse from re-infection in TB recurrence.

Implications of available evidence. We propose that critical transmission events may be missed if within-host diversity is overlooked. Relying on standard sequencing depth can therefore have a major impact on whether we accurately identify transmission clusters and corresponding risk factors for transmission. We suggest – at a minimum – targeted deep sequencing of highly contagious cases, in addition to routine sequencing, to more accurately resolve TB transmission dynamics and help guide appropriate public health interventions.

Abstract – 250/250 words

Background. Tuberculosis disproportionately affects the Canadian Inuit population. To address this problem, it is imperative we understand transmission dynamics. Here, we investigate whether 'deep' sequencing provides additional resolution compared to standard sequencing, with implications for TB control in high-transmission environments.

Methods. 48/50 samples from an outbreak (2011-2012), and 14/15 samples from the preceding five years from an Arctic village, were deep sequenced (to ~500-1000x) using Illumina HiSeq. Reads were aligned to a novel local reference genome, generated using PacBio SMRT sequencing. Consensus single nucleotide polymorphisms (cSNPs) and heterogeneous alleles (hSNPs, where the alternative allele was in > 5% and < 95% of reads) were identified. Core cSNPs were used to calculate genetic distances, and hSNPs were reviewed between samples.

Findings. In contrast with previous genomic analyses using ~50x depth, deep sequencing allowed us to identify a novel super-spreader. This case, previously thought to have transmitted to four others, likely transmitted to up to 17 other cases during the outbreak (35% of all cases that year). 76% attended the same community gathering houses as the putative source, supporting a major role for these social venues in the outbreak. hSNP analysis also helped confirm other transmission events, and discriminate relapse from re-infection in TB recurrence.

Interpretation. Deep sequencing and investigation of within-host diversity may be critical to accurately identify super-spreaders and corresponding transmission clusters. We propose – at a minimum - targeted deep sequencing of highly contagious cases, in addition to routine sequencing, to facilitate public health investigations.

Funding. National Institutes of Health.

Introduction.

Tuberculosis (TB) in Canada is highest among the Inuit, an Indigenous population with a rate over 300x that of the non-Indigenous Canadian-born population in 2016. [1] Canada recently set a goal of TB elimination in the Inuit by 2030,[1] which will not be achieved without halting ongoing transmission. Previous studies have used genomic data either alone or in conjunction with classical epidemiology to investigate TB transmission dynamics in the Canadian North,[2-4] with the aim of identifying clusters to help guide public health interventions. Thus far, such studies have relied on identifying consensus single nucleotide polymorphisms (cSNPs), consistent with prevailing methodology in this field.

Recent studies suggest that incorporation of within-host diversity into genomic analyses may provide greater resolution of transmission than cSNP-based approaches alone.[5,6] This may be particularly important for investigation of outbreaks occurring over short time scales and/or in settings such as the Canadian North, where the genetic diversity of circulating strains is especially low. In both of these circumstances, it is common to find many samples separated by zero cSNPs, hindering accurate source ascertainment. To investigate this hypothesis, we used deep sequencing (i.e., to ~10-fold more than standard, or 500-1000x) to re-evaluate transmission in a densely-sampled outbreak in Nunavik, Québec.

This outbreak, which has been previously described,[4,7] comprised 50 microbiologically-confirmed cases of TB who were diagnosed in a single Inuit community between 2011-2012 - a rate of 5,359/100,000 for that year. Genomic epidemiology analyses using sequencing depths of ~50x that are standard in such work, identified multiple clusters of transmission in this outbreak,

[4] however, there was insufficient genetic variation detected to infer precise person-to-person transmission events within these subgroups. In this study, we illustrate how within-host diversity can be incorporated into transmission analyses and in doing so, find new features of the transmission networks and dynamics during the outbreak, in particular, identifying a previously unrecognized superspreading event. We highlight a role for deep sequencing in public health investigations, with implications for TB control in Canada's North as well as other high-transmission environments.

Materials and methods.

Study subjects. All 50 samples from the 2011-2012 outbreak were included, as well as samples from all cases (n=15) diagnosed in same village in the preceding five years (2007 onwards), 13/15 of which were caused by the same strain of *M. tuberculosis* (the 'Major [Mj]-III' sublineage [3]). All cases were Lineage 4 (Euro-American [4]) and all had pulmonary TB. Cross-contamination between samples was ruled out as previously described.[4]

DNA extraction and sequencing. Laboratory methods are described in detail in the **Supplementary Material**. The Illumina HiSeq 4000 was used for paired-end 100bp sequencing. To obtain the target depth of coverage, pooled libraries were run on four independent lanes.

Bioinformatics. Quality control of genomic data is described in detail in the **Supplemental Material.** Reads were aligned using the Burrows Wheeler Aligner MEM algorithm (v.0.7.15 [8]) to the H37Rv reference (NC_000962.3 in the National Center for Biotechnology Information [NCBI] RefSeq database) and sorted using Samtools (v.1.5 [9]). Analyses were later repeated using

125 a local reference genome (described below). Reads with ambiguous mapping were excluded, as
126 were reads with excessive soft-clipping (i.e., more than 20% of read length) based on our
127 previous work. [6] Duplicate reads were marked using Picard MarkDuplicates (v.2.9.0,
128 https://broadinstitute.github.io/picard/) and reads were locally re-aligned around indels using
129 Genome Analysis ToolKit (GATK, v.3.8 [10]). All sites were called using GATK's Unified
130 Genotyper algorithm.
131
132 Variants were filtered for quality using custom Python scripts (v.3.6) with the following
133 thresholds: Phred < 50, Root Mean Squared Mapping Quality (RMS-MQ) ≤ 30, depth (DP) < 20,
134 Fisher Strand Bias (FS) ≥ 60 and read position strand bias (ReadPos) < -8.[6,11] cSNPs were
135 classified as positions where ≥ 95% of reads were the alternative allele (ALT), hSNPs were
136 classified as positions where > 5% and < 95% of reads were ALT, and positions with the ALT
137 present in < 5% of reads were classified as 'reference'. We also compared inferences of
138 transmission from this analysis to i) when these thresholds were increased to the minimum
139 values among cSNPs in the initial H37Rv analysis, and ii) when cSNPs were classified using a
140 threshold of ≥ 99%, and hSNPs were classified when 1% < ALT < 99%, in order to assess the
141 robustness of inferences to different filtering protocols.
142
143 Low-quality variants, variants in proline-proline-glutamic acid (PE) and proline-glutamic-
144 acid/polymorphic-guanine-cytosine-rich sequence (PE_PGRS) genes, transposons, phage and
145 integrase, and positions with missing data, were excluded. All samples were drug-susceptible,
146 except for MT-6429, which was rendered resistant to isoniazid by a frameshift deletion at

147 position 1284 in the catalase-peroxidase gene *katG*. As such, positions associated with drug
148 resistance were not masked in this analysis.
149
150 Concatenated cSNP alignments were generated excluding positions with hSNPs. Pairwise cSNP
151 distances between samples were computed using snp-dists (v.0.6, available at
152 https://github.com/tseemann/snp-dists). The frequency of hSNPs at each position in the genome
153 was tabulated and hSNPs were reviewed to identify variants shared between samples.
154
155 ***Phylogenetics and clustering.*** Core cSNP alignments were used to generate maximum
156 likelihood trees using IQ-Tree (v.1.6.8 [12]). Model selection was based on the lowest Bayesian
157 Information Criterion. Hierarchical Bayesian Analysis of Population Structure [13] was run in R
158 (v.3.5.2) to identify clusters. See the **Supplementary Material** for additional detail.
159
160 ***Single Molecule Real-Time (SMRT) sequencing and assembly.*** To examine the influence of
161 potential alignment errors in identification of hSNPs, we used SMRT sequencing with the
162 PacBio RSII platform to create a local reference genome for the outbreak. Sample MT-0080 was
163 chosen for sequencing because this was previously identified as the probable source for as many
164 as 19 of the 50 cases diagnosed in 2011-2012.[4] A single colony from the culture was selected for
165 SMRT sequencing and Illumina MiSeq (for polishing of the long-read assembly). Further detail
166 is provided in the **Supplementary Material**.
167

168 Long-reads were assembled and corrected using Canu (v.1.7.1 [14]). Pilon (v.1.23 [15]) was then used
169 to polish the assembly and was re-run until no further corrections were possible. Quast (v.5.0.2,
170 [16]) was used to evaluate assembly quality and RASTtk (v.2.0 [17]) was used for annotation.
171
172 ***Epidemiological data.*** Detailed epidemiological and clinical data were collected by the local
173 healthcare team and the regional public health unit, as previously described. [4,7]
174
175 ***Statistical analyses.*** Differences in proportions were assessed using a two-sample test of
176 proportions. Distributions were compared using the Wilcoxon-Mann-Whitney test. Statistics
177 were calculated using Stata (v.15, StataCorp, College Station, TX, USA).
178
179 ***Data availability.*** Illumina HiSeq and PacBio SMRT sequencing data are available on the
180 NCBI's Sequence Read Archive under BioProject PRJNA549270, along with the complete
181 assembly for MT-0080.
182
183 ***Ethics***. Ethics approval was obtained from the Institutional Review Board (IRB) of the Harvard
184 T.H. Chan School of Public Health (IRB18-0552) and the IRB of McGill University Faculty of
185 Medicine (IRB A02-M08-18A). All data was analyzed in non-nominal fashion. This study was
186 done with approval of and in collaboration with the Nunavik Regional Board of Health and
187 Social Services.
188

Role of the funding source. The funding organizations had no role in study design, data collection, analysis or interpretation of data, nor in the writing of this manuscript and decision to submit the paper for publication.

Results.

62/65 (95·4%) available TB samples from cases diagnosed between 2007-2012 were successfully sequenced and passed quality control. This included 48/49 (98·0%) of the samples with an identical Mycobacterial Interspersed Repetitive Units Variable Number Tandem Repeats (MIRU-VNTR) pattern during the outbreak year. The remaining three samples could not be re-grown.

Average genome coverage and depth across the H37Rv reference was 98·64% [SD 0·07%] and 714·53 [SD 92·68], respectively. Our primary filtering protocol yielded 51,430 cSNPs and 4,897 hSNPs across all individual samples (**Table S2**). Excluding positions that were invariant compared to the reference or where any sample was missing and/or was low-quality resulted in a core alignment of 860 cSNP positions and 136 hSNP positions (note, these are not mutually exclusive, as positions with cSNPs in some samples may have hSNPs in others).

42 positions had hSNPs that were shared across all 62 samples (**Table 1**, **Supplementary Dataset 1A**). Depth of coverage at these positions was, on average, 39% higher than the average depth across the same sample (SD 36·7%, **Supplementary Dataset 1B**). Along with manual review of alignments (**Figure S1)**, this suggested that many of these were false positives,

211 potentially due to alignment error (e.g., from underlying structural variation in our samples
212 compared to the H37Rv reference).
213
214 To address this, we generated a local reference genome for the outbreak, MT-0080_PB. Quality
215 metrics for the MT-0080_PB assembly are given in **Table S1**. Compared to H37Rv, mean
216 genome coverage and depth were higher with MT-0080_PB (at 99·33% [SD 0·09%] and 717·07
217 [SD 93·01], respectively), fewer positions were missing/low-quality ($p < 0·00005$, **Table 1**), and
218 overall, fewer variable positions were detected ($p < 0·00005$). While core cSNP distances were
219 similar between samples regardless of the reference (**Table 1**), the number of hSNPs was greatly
220 reduced using MT-0080_PB (**Table S2**); while 4,897 hSNPs were identified across all individual
221 samples using H37Rv, only 134 hSNPs were identified using MT-0080_PB. There were also no
222 hSNPs shared across all 62 samples using MT-0080_PB. Together, these findings support our
223 hypothesis that alignment error is responsible for many of the detected variants, and indicate a
224 local reference is important for accurate identification of hSNPs. All further results presented are
225 based on the MT-0080_PB alignment.
226
227 A maximum likelihood tree was generated from 103 core cSNP positions compared to MT-0080
228 (**Figure 1A**). Consistent with previous work,[4] hierBAPS identified two main sub-lineages ('Mj-
229 V' and 'Mj-III' per [3]), with three sub-clusters (Mj-IIIA/B/C).
230
231 ***hSNPs identify super-spreaders and more accurately resolve transmission clusters.*** The core
232 cSNPs and hSNP alignment is shown in **Supplemental Dataset 2A**. Overlaying hSNPs with the
233 cSNP-based analysis revealed a novel super-spreader (MT-504) in Cluster Mj-IIIB, undetected

234 by genomic analyses relying on lower sequencing depth. MT-504 had smear-positive cavitary
235 disease and was diagnosed in late 2011; previous analyses had found this case shared a single
236 cSNP with four other cases diagnosed from March – December of 2012 (position 276,685
237 according to H37Rv / 276,544 in the MT-0080_PB alignment, **Supplementary Dataset 2**).
238 Coupled with epidemiological data on contact (shared attendance at local community 'gathering
239 houses', social venues specifically identified by public health during the outbreak), this strongly
240 supported transmission from MT-504 to other members of this subgroup. In contrast, the other
241 subgroup of Mj-IIIB with 13 cases did not share this cSNP, initially refuting transmission in this
242 low-diversity setting (**Figure 1B**). Instead, it was previously postulated that the first smear-
243 positive case in this subgroup (MT-2474, diagnosed in May 2012) led to the majority of
244 transmission (note, the first smear-negative case in this subgroup was diagnosed in March 2012).
245 However, deep sequencing data suggest otherwise; these data show that MT-504 harboured both
246 the reference (133 reads [19·1%]) allele, present in the subgroup of 13, as well as the alternative
247 allele (563 reads [80·9%]) at this position (**Figure 1C**). As MT-504 was the first contagious case
248 diagnosed in Mj-IIIB, and all 13 cases in this subgroup had attended or resided in a gathering
249 house (with 9/13 [69·2%] reporting attendance at the same houses as MT-504), this strongly
250 suggests that MT-504 is in fact the most probable source for both subgroups.
251
252 ***hSNP analysis adds support for suspected transmission.*** Sample 68995 and MT-5543 were the
253 only two strains from the Mj-VA sub-lineage in this village and were separated by zero core
254 cSNPs. As previous analysis indicated Mj-VA strains from other villages were distantly related,[4]
255 this had suggested direct transmission between these historical cases. This hypothesis is strongly
256 supported by hSNP analysis, as the samples share hSNPs that are not found in any other sample

in the dataset. This held true with highly conservative filtering thresholds used **(Supplemental Dataset 2B**). Importantly, these hSNPs were not detected when using H37Rv as the reference.

Potential utility for discriminating TB recurrence. During the outbreak year, six individuals had TB recurrence, wherein they had previously been treated (and cured) using Directly Observed Therapy for active TB. Paired samples were available for two of these individuals (Patient 1: samples MT-5195 in 2007 and MT-1838 in 2012; Patient 2: samples MT-5543 in 2007 and MT-1206 in 2012, **Figure 1A**). cSNP-based analyses suggested their second episodes of TB were due to re-infection, rather than relapse; investigation of within-host diversity strongly supported this conclusion, as there was neither evidence of mixed strains at baseline nor super-infection at their second episode (**Supplemental Dataset 2A/B/C**).

Impact of altering cSNP and hSNP thresholds. To ensure we were not missing lower frequency variants using the prior cSNP/hSNP thresholds, we re-ran our analysis such that hSNPs were classified when 1% < ALT < 99%. Quality scores for individual cSNPs and hSNPs are given in **Table S3** and the core cSNP/hSNP alignment is shown in **Supplemental Dataset 2C**. While our primary analysis using a threshold of > 95% for cSNPs identified a single cSNP (A>G) shared across all samples compared to MT-0080_PB, close examination of the MT-0080 deep sequencing data (obtained using DNA from a sweep of the plate) showed that this sample had both alleles at this position, with only the minority 'A' allele (33 reads/1189 [2·8%]) isolated for SMRT sequencing. Based on this, we recommend sequencing samples both using a clean sweep (with an alternative sequencing platform) and a single colony pick when generating a reference genome for TB, as using the latter alone may introduce error and affect epidemiological

inferences. With this exception, no other informative hSNPs were detected using these thresholds.

Discussion.

As the TB epidemic continues among the Canadian Inuit, targeted public health interventions are essential to halt ongoing transmission. In order to do so, it is important that transmission events and associated risk factors are accurately identified. Our previous work suggested that hSNP analysis could enhance resolution of TB transmission [6]. To investigate how this approach could be applied for TB control, we used deep sequencing to re-examine a major TB outbreak in the Canadian Arctic.

In contrast to previous genomic analyses,[3,4] using deep sequencing allowed us to identify a novel super-spreader who was likely responsible for nearly 1/3 of the cases from 2011-2012. Super-spreading has been described in a number of pathogens,[18] including TB.[19] Our findings suggest this can play an important role in driving TB outbreaks. We therefore propose that investigation of within-host diversity is necessary to ensure detection of such super-spreaders in this context, and potentially other high-transmission environments as well. We typically only know that a person is a super-spreader retrospectively, i.e., once they have already transmitted to many others and once corresponding genomic data is available. However, if we can identify the characteristics associated with this phenomenon at the population level, this could be used prospectively to predict whether cases are likely to be super-spreaders - as they are diagnosed. This could allow resources to be better allocated, for example, investigation of their contacts could be prioritized and/or targeted screening of the social venues they attend could be rapidly

initiated, potentially leading to faster detection of secondary cases and initiation of prophylaxis for new infections. In the case of MT-504, nearly all of the secondary cases had attended the same local community gathering houses as the putative source; this strongly suggests the importance of these venues in facilitating transmission in this setting.

Several studies have used genomics to investigate TB recurrence,[20-22] however, the methods used to assess for mixed infection at either time point have been inconsistent and may not be sufficient to discriminate recurrence in settings with low strain diversity. In this analysis, we demonstrate proof-of-principle that deep sequencing can help evaluate mixed infection and, in doing so, help rule out relapse. The distinction between relapse and re-infection is important at individual and population levels; high rates of relapse in a community would indicate a problem with treatment or adherence, potentially warranting changes to clinical management, while re-infection would indicate the need for public health interventions such as activate case finding. Also, individuals in Nunavik who have had prior treatment for active TB disease in the past are also not routinely offered prophylaxis on re-exposure, based on historical data suggesting ~80% protection is afforded by prior infection.[23] The degree to which re-infection drives recurrence in Nunavik is currently unknown, but if re-infection is the primary cause, this clinical practice may need to be re-evaluated. Population-level genomic studies are urgently needed to evaluate this.

To use deep sequencing to investigate within-host diversity, it is critical we minimize false positive hSNPs. We have shown that using a local reference strain not only reduces error, but improves detection of epidemiologically-informative variants. O'Toole *et al*[24]. previously suggested clinical TB strains are needed to detect key virulence genes with reference-based

325 analyses. We propose these are also warranted for hSNP analysis. Where possible, we
326 recommend using long-read sequencing to generate complete and local reference genomes.
327
328 Overall, our study has a number of strengths. Firstly, we had access to a densely-sampled
329 outbreak, which was previously investigated using 'standard' sequencing depth and for which
330 detailed epidemiological data was available. This allowed us to readily compare methodological
331 approaches, showing how and when deep sequencing might be beneficial for public health. In
332 doing this, we have identified important methodological considerations for hSNP detection, with
333 implications for transmission analyses, but also potentially for resistance prediction as well.[25]
334 Finally, the use of long-read data has allowed us to completely assemble a novel TB genome
335 from Nunavik. We hope that this will serve as a valuable tool for future studies of transmission
336 in Nunavik as well as other Inuit territories.
337
338 A potential limitation of this work is that, given the historical nature of the outbreak, deep
339 sequencing was done using DNA extracted from culture. Due to methodological challenges of
340 sequencing directly from sputum,[26-28] few studies have examined the effect of culture on genome
341 diversity. A recent study by Shockey *et al*[29]. using cSNPs suggests that variation may be lost
342 during the culturing process, however authors did not examine the impact on hSNPs or
343 transmission, and several other studies [27,28,30] previously found congruent results between cSNP
344 analyses from culture versus raw samples. In terms of hSNPs, Votintseva *et al*. found no
345 difference in the number detected between approaches.[27] While [28,30] reported detecting fewer
346 hSNPs with sequencing from culture versus from sputum, in [30], the median hSNPs was only 4.5
347 versus 5 hSNPs, respectively – a difference that may not be clinically significant, regardless of

statistical significance. Given the inconsistency of results and paucity of data, further study is needed to understand how hSNP diversity may be affected by the culturing process, and to assess whether this affects inferences of transmission. We note that it is likely that enhanced detection of the hSNPs present in sputum would improve the resolution over that which we present in this work.

Another potential limitation is that, while we can compare the epidemiological inferences made between our previous analysis and our deep sequencing analysis, the bioinformatics pipelines themselves are not directly comparable. Methods to accurately identify hSNPs are currently an area of active research. We illustrated in our recent paper [6] that additional steps and strict thresholds must be used to minimize false positive hSNPs, and have conducted the current analysis in consideration of this. However, we note that pre-filtering, our 2015 analysis found that MT-504 had only five reference alleles at position 276,685 in the H37Rv alignment (out of 75) and randomly downsampling the current data to simulate ~50x yielded similar results (5/47 reads at position 276,544 aligned to MT-0080_PB). Given that most genomic studies of TB employ thresholds of 75-90% allele frequency to classify cSNPs at standard sequencing depth, our inferences regarding transmission would remain largely unchanged, reinforcing the need for greater depth to make such critical distinctions.

In summary, we have found evidence of heterogeneous variants with important epidemiologic implications that would not have been detected with standard methods and filtering criteria. This illustrates that genomic methods, while powerful, still require careful interpretation and can still harbor considerable ambiguity when comparing very close links in a transmission chain. We

371 demonstrate that deep sequencing can aid in TB transmission analyses, in particular by allowing
372 accurate identification of super-spreaders and key associated epidemiological characteristics.
373 This work has important implications for TB control in the Canadian North as well as other
374 regions with high TB transmission; as next-generation sequencing becomes a mainstay in public
375 health surveillance, it is critical we recognize the limitations of analyses done using routine
376 sequencing data and consider where and when deep sequencing might be warranted. Although
377 costs continue to decline, we recognize deep sequencing of all samples in an outbreak may not
378 be economically viable for every public health unit. As such, we propose that public health units
379 using routine sequencing for tuberculosis also incorporate – at a minimum – targeted deep
380 sequencing of the more contagious (e.g., smear-positive) cases to ensure accurate identification
381 of super-spreaders and clusters of transmission. This will ultimately help TB control programmes
382 better understand the risk factors for such transmission and enable prioritization of public health
383 resources in future outbreaks.
384
385 **Declaration of Interests.** Authors have no conflicts of interest to declare.
386
387 **Contributions.** RSL and WPH conceived and designed the study. RSL designed and did the
388 analyses, made the tables and figures, interpreted the data, and wrote the first draft of
389 manuscript. JFP provided epidemiological data. MAB provided the bacterial samples, as well as
390 laboratory reagents, Biosafety Level 3 access and technician labour in-kind to RSL. FM did the
391 culture and DNA extraction for the HiSeq and PacBio SMRT sequencing. WPH reviewed the
392 initial draft. All authors provided feedback and reviewed the final version of the manuscript

before submission. RSL had full access to the data and final responsibility for the decision to submit for publication.

Funding. This work was supported by an R01 grant from the National Institutes of Health, awarded to WPH (R01AI128344). RSL is also supported by a Fellowship from the Canadian Institutes of Health Research (MFE 152448). MAB holds a Canadian Institutes of Health Research Foundation Grant (FDN-148362).

Acknowledgements. We would like to thank the council of the village for their ongoing support of this work. We also acknowledge staff from the Centre Locales des Services Communautaires and the Nunavik Regional Board of Health and Social Services for their hard work during the outbreak. We would like to thank Dr. Hafid Soualhine (currently at the National Reference Centre for Mycobacteriology in the National Microbiology Laboratory, Public Health Agency of Canada) for previous confirmation of the frameshift deletion in *katG* of MT-6429. We also thank Dr. Anders Gonçalves da Silva and Dr. Glen Carter of the Microbiological Diagnostic Unit Public Health Laboratory at the University of Melbourne for their helpful input on high-quality SMRT sequencing. Library preparation and sequencing for all samples was done at the Genome Québec / McGill Innovation Centre and high-performance computing was done using the Odyssey cluster from the Faculty of Arts and Science, Harvard University.

414 **References.**

415 1. Kanatami IT. Inuit Tuberculosis Elimination Framework. 2018.

416 2. Tyler AD, Randell E, Baikie M, et al. Application of whole genome sequence analysis to
417 the study of Mycobacterium tuberculosis in Nunavut, Canada. *PLoS One* 2017; **12**(10):
418 e0185656.

419 3. Lee RS, Radomski N, Proulx JF, et al. Population genomics of Mycobacterium
420 tuberculosis in the Inuit. *Proc Natl Acad Sci U S A* 2015; **112**(44): 13609-14.

421 4. Lee RS, Radomski N, Proulx JF, et al. Reemergence and amplification of tuberculosis in
422 the Canadian arctic. *J Infect Dis* 2015; **211**(12): 1905-14.

423 5. Worby CJ, Lipsitch M, Hanage WP. Shared Genomic Variants: Identification of
424 Transmission Routes Using Pathogen Deep-Sequence Data. *Am J Epidemiol* 2017; **186**(10):
425 1209-16.

426 6. Martin MA, Lee RS, Cowley LA, Gardy JL, Hanage WP. Within-host Mycobacterium
427 tuberculosis diversity and its utility for inferences of transmission. *Microb Genom* 2018; **4**(10).

428 7. Lee RS, Proulx JF, Menzies D, Behr MA. Progression to tuberculosis disease increases
429 with multiple exposures. *Eur Respir J* 2016; **48**(6): 1682-9.

430 8. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
431 *arXiv* 2013: 1303.3997v2.

432 9. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and
433 SAMtools. *Bioinformatics* 2009; **25**(16): 2078-9.

434 10. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce
435 framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010; **20**(9): 1297-
436 303.

437 11. Lee RS, Behr MA. Does Choice Matter? Reference-Based Alignment for Molecular
438 Epidemiology of Tuberculosis. *J Clin Microbiol* 2016; **54**(7): 1891-5.
439 12. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective
440 stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 2015;
441 **32**(1): 268-74.
442 13. Cheng L, Connor TR, Siren J, Aanensen DM, Corander J. Hierarchical and spatially
443 explicit clustering of DNA sequences with BAPS software. *Mol Biol Evol* 2013; **30**(5): 1224-8.
444 14. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable
445 and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome*
446 *Res* 2017; **27**(5): 722-36.
447 15. Walker BJ, Abeel T, Shea T, et al. Pilon: an integrated tool for comprehensive microbial
448 variant detection and genome assembly improvement. *PLoS One* 2014; **9**(11): e112963.
449 16. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for
450 genome assemblies. *Bioinformatics* 2013; **29**(8): 1072-5.
451 17. Brettin T, Davis JJ, Disz T, et al. RASTtk: a modular and extensible implementation of
452 the RAST algorithm for building custom annotation pipelines and annotating batches of
453 genomes. *Sci Rep* 2015; **5**: 8365.
454 18. Stein RA. Super-spreaders in infectious diseases. *Int J Infect Dis* 2011; **15**(8): e510-3.
455 19. Kline SE, Hedemark, L.L., Davies, S.F. Outbreak of tuberculosis among regular patrons
456 of a neighborhood bar. *New Engl J Med* 1995; **333**(4): 222-7.
457 20. Witney AA, Bateson AL, Jindani A, et al. Use of whole-genome sequencing to
458 distinguish relapse from reinfection in a completed tuberculosis clinical trial. *BMC Med* 2017;
459 **15**(1): 71.

460 21. Bryant JM, Harris SR, Parkhill J, et al. Whole-genome sequencing to establish relapse or
461 re-infection with Mycobacterium tuberculosis: a retrospective observational study. *The Lancet*
462 *Respiratory Medicine* 2013; **1**(10): 786-92.
463 22. Guerra-Assuncao JA, Houben RM, Crampin AC, et al. Recurrence due to relapse or
464 reinfection with Mycobacterium tuberculosis: a whole-genome sequencing approach in a large,
465 population-based cohort with a high HIV infection prevalence and active follow-up. *J Infect Dis*
466 2015; **211**(7): 1154-63.
467 23. Menzies D. Issues in the management of contacts of patients with active pulmonary
468 tuberculosis. *Can J Public Health* 1997; **3**: 197-201.
469 24. O'Toole RF, Gautam SS. Limitations of the Mycobacterium tuberculosis reference
470 genome H37Rv in the detection of virulence-related loci. *Genomics* 2017; **109**(5-6): 471-4.
471 25. Liu Q, Via LE, Luo T, et al. Within patient microevolution of Mycobacterium
472 tuberculosis correlates with heterogeneous responses to treatment. *Sci Rep* 2015; **5**: 17507.
473 26. Brown AC, Bryant JM, Einer-Jensen K, et al. Rapid Whole-Genome Sequencing of
474 Mycobacterium tuberculosis Isolates Directly from Clinical Samples. *J Clin Microbiol* 2015;
475 **53**(7): 2230-7.
476 27. Votintseva AA, Bradley, P., Pankhurst, L., del Ojo Elias, C., Loose, M., Nilgiriwala, K.,
477 Chatterjee, A., Smith, E.G., Sanderson, N., Walker, T.M., Morgan, M.R., Wyllie, D.H., Walker,
478 A.S., Peto, T.E.A., Crook, D.W., Iqbal, Z. Same-day diagnostic and surveillance data for
479 tuberculosis via whole genome sequencing of direct respiratory samples. *J Clin Microbiol* 2017.
480 28. Doyle RM, Burgess, C., Williams, R., Gorton, R., Booth, H., Brown, J., Bryant, J.M.,
481 Chan, J., Creer, D., Holdstock, J., Kunst, H., Lozewicz, S., Platt, G., Yara Romero, E., Speight,
482 G., Tiberi, S., Abubakar, I., Lipman, M., McHugh, T.D., Breuer, J. Direct Whole-Genome

483 Sequencing of Sputum Accurately Identifies Drug-Resistant Mycobacterium tuberculosis Faster
484 than MGIT Culture Sequencing. *J Clin Microbiol* 2018; **56**(8): e00666-18.
485 29. Shockey AC, Dabney J, Pepperell CS. Effects of Host, Sample, and in vitro Culture on
486 Genomic Diversity of Pathogenic Mycobacteria. *Frontiers in Genetics* 2019; **10**.
487 30. Nimmo C, Shaw LP, Doyle R, et al. Whole genome sequencing Mycobacterium
488 tuberculosis directly from sputum identifies more genetic diversity than sequencing from culture.
489 *BMC Genomics* 2019; **20**(1): 389.
490

491 **Tables**

492

493 **Table 1.** Comparison of alignments to H37Rv and MT-0080_PB

494

	H37Rv (4,411,532 bp)	MT-0080_PB (4,426,525 bp)	P value
Number of positions according to reference genome			
Invariant reference across all samples, n (%)	4,018,786 (91·10%)	4,267,097 (96·40%)	1·0 [a]
Position was missing / low quality in at least one sample, n (%)	391,761 (8·88%)	159,259 (3·60%)	< 0·00005 [a]
Position was an c/hSNP in at least one sample, n (%)	985 (0·22%)	169 (0·00%)	< 0·00005 [a]
Shared cSNPs across all samples, n (%)	764 (0·02%)	1 (0·00%)	< 0·00005 [a]
Shared hSNPs across all samples, n (%)	42 (0·00%)	0 (0%)	< 0·00005 [a]
Core pairwise distances			
Core cSNPs vs. reference, median (range)	791 (790-792)	3 (1-76)	0·0001 [b]
Core cSNPs between samples, median (range)	3 (0-64)	3 (0-77)	0·0658 [b]

495 Legend. Based on these filters: Phred < 50, Root Mean Square Mapping Quality (RMS-MQ) ≤ 30, depth (DP) < 20, Fisher Strand Bias (FS) ≥ 60 and read
496 position strand bias (ReadPos) < -8 and an allelic fraction of 95% for cSNPs, with hSNPs classified when 5% < ALT < 95%. Quality metrics for the individual
497 cSNPs/hSNPs identified in each sample are given in **Table S2**. [a] Two sample test for difference in proportions. [b] Wilcoxin-Mann-Whitney test.

Figure Legends.

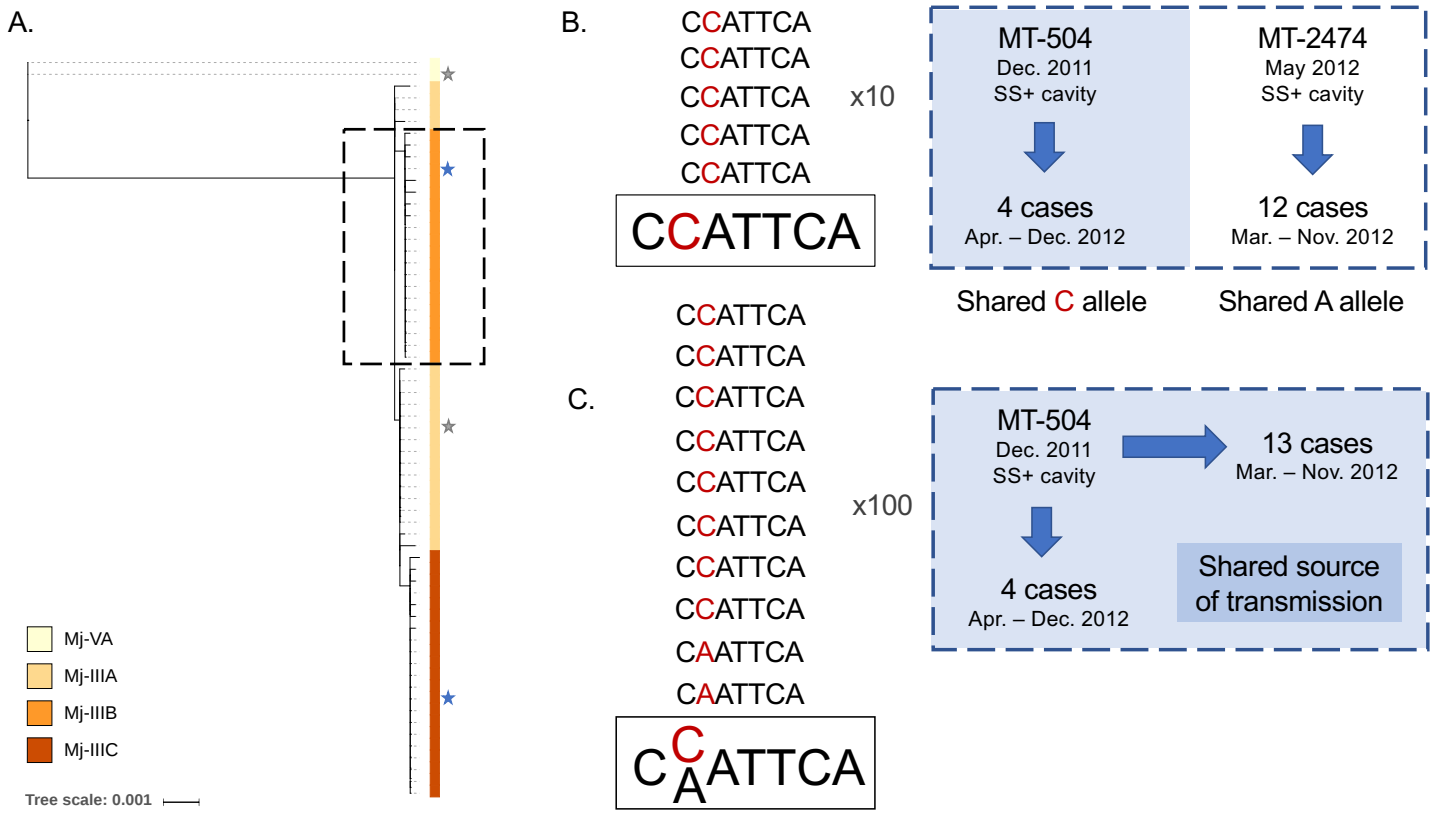
Figure 1. Transmission of *M. tuberculosis* in village K.

Panel A. Maximum likelihood tree of 62/65 cases diagnosed between 2007-2012 in village K based on consensus single nucleotide polymorphisms (cSNPs). After aligning to a local reference, MT-0080_PB, cSNPs were identified based on a minimum threshold of ≥ 95% of reads supporting the alternative allele. A core cSNP alignment was then produced with 103 positions. IQ-Tree (v.1.6.8 [12]) was then used to generate the tree using a KP3 model with correction for ascertainment bias. Model selection was based on the lowest Bayesian Information Criterion. Clusters were identified using hierarchical Bayesian Analysis of Population Structure.[13] These clusters were consistent with the sub-lineages previously identified in [3,4], thus only sub-lineage names are indicated. During this time period, there were two individuals who had a second episode of TB; stars are used to highlight these samples, with a different colour for each patient. MT-0080 is included in the alignment as the deep sequencing data from a sweep of all colonies identified a cSNP compared to the MT-0080_PB reference, which itself was generated from a single colony pick.

Panel B. Standard sequencing (to ~40-50x), along with epidemiological data, had indicated that the Major [Mj]-IIIB sub-lineage was comprised of two subgroups of five and 13 patients, respectively.[4] MT-504 was the suspected source case for the subgroup of five, which all shared a 'C' allele at position 276,685 in H37Rv (position 276,544 in MT-0080_PB). In contrast, all members of the subgroup of 13 shared an 'A' at this position. Previously, MT-2474 was the suspected source case for this subgroup; this case was the first person with smear-positive (SS+) cavitary disease diagnosed in this subgroup.

Panel C. In contrast to standard sequencing, deep sequencing data revealed that, in fact, MT-504 – the presumed source for the subgroup of five cases and the first highly contagious case diagnosed in Mj-IIIB during the outbreak year – had both 'C' and 'A' alleles at this position (563:133 of reads, respectively), suggesting this was in fact the most probable source for both subgroups.

Figure 1.


A.
Mj-VA
Mj-IIIA
Mj-IIIB
Mj-IIIC
Tree scale: 0.001
B.
CCATTCA
CCATTCA
CCATTCA
CCATTCA
CCATTCA
x10
CCATTCA
MT-504
Dec. 2011
SS+ cavity
4 cases
Apr. – Dec. 2012
MT-2474
May 2012
SS+ cavity
12 cases
Mar. – Nov. 2012
Shared C allele
Shared A allele
C.
CCATTCA
CCATTCA
CCATTCA
CCATTCA
CCATTCA
CCATTCA
CCATTCA
CCATTCA
CAATTCA
CAATTCA
x100
CCATTCA
MT-504
Dec. 2011
SS+ cavity
13 cases
Mar. – Nov. 2012
4 cases
Apr. – Dec. 2012
Shared source of transmission