Reemergence and Amplification of Tuberculosis in the Canadian Arctic

Robyn S. Lee,1,4,5,a Nicolas Radomski,5,a Jean-Francois Proulx,6 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. Behr4,5

Departments of 1Epidemiology, Biostatistics and Occupational Health, 2Medicine, and 3Human Genetics, McGill University, 4McGill International TB Centre, 5The Research Institute of the McGill University Health Centre, 6Mycobacteriology Laboratory, Royal Victoria Hospital, and 7Respiratory Epidemiology and Clinical Research Unit, Montreal Chest Institute, McGill University Health Centre, 8Nunavik Regional Board of Health and Social Services, Kuujjuaq, and 9Laboratoire 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, IIB, and IIIC. Between 2007 and 2010, cluster IIIA predominated (11 of 22 vs 2 of 40; P < .001), whereas in 2011–2012 clusters IIB (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 Mycobacterium 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.

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 km² 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 log₁₀ Perror). 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 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.

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
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 \textit{fadE}4) 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 $\geq 2$ 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 tentative source reported symptoms for 4 months before diagnosis,
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*, *Rv1833c*, and *Rv2263* (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.

<table>
<thead>
<tr>
<th>Position in H37Rv</th>
<th>Gene code</th>
<th>Gene name</th>
<th>Reference allele</th>
<th>Alternative allele</th>
<th>Mutation</th>
<th>HIIA</th>
<th>IIB</th>
<th>IIIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1331500</td>
<td>Rv1188</td>
<td>A</td>
<td>G</td>
<td>Nonsynonymous</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>1507283</td>
<td>Rv1392</td>
<td>netK</td>
<td>C</td>
<td>Synonymous</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>2025484</td>
<td>Rv1784</td>
<td>C</td>
<td>G</td>
<td>Nonsynonymous</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>22398</td>
<td>Rv0018c</td>
<td>C</td>
<td>G</td>
<td>Nonsynonymous</td>
<td>G</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>42301</td>
<td>N/A</td>
<td>G</td>
<td>A</td>
<td>Intergenic</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>383629</td>
<td>Rv0315</td>
<td>G</td>
<td>A</td>
<td>Nonsynonymous</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>428744</td>
<td>Rv0355c</td>
<td>PPE8</td>
<td>A</td>
<td>Nonsynonymous</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>55829</td>
<td>N/A</td>
<td>G</td>
<td>A</td>
<td>Intergenic</td>
<td>G</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>1011972</td>
<td>Rv0908</td>
<td>ctgE</td>
<td>G</td>
<td>Nonsynonymous</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>2115139</td>
<td>Rv1866</td>
<td>G</td>
<td>A</td>
<td>Nonsynonymous</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>3251536</td>
<td>Rv2392</td>
<td>C</td>
<td>T</td>
<td>Synonymous</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>3280352</td>
<td>Rv2904c</td>
<td>C</td>
<td>A</td>
<td>Nonsynonymous</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>3307489</td>
<td>Rv3090c</td>
<td>T</td>
<td>C</td>
<td>Synonymous</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>3569065</td>
<td>Rv3010c</td>
<td>T</td>
<td>C</td>
<td>Nonsynonymous</td>
<td>G</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>3910540</td>
<td>Rv3492c</td>
<td>G</td>
<td>A</td>
<td>Synonymous</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>4939885</td>
<td>Rv3652</td>
<td>PE_FG856</td>
<td>C</td>
<td>Nonsynonymous</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>30642</td>
<td>Rv0026</td>
<td>C</td>
<td>T</td>
<td>Synonymous</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>403587</td>
<td>Rv0338c</td>
<td>C</td>
<td>T</td>
<td>Synonymous</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>571113</td>
<td>N/A</td>
<td>C</td>
<td>A</td>
<td>Intergenic</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>773818</td>
<td>Rv0675</td>
<td>ecbA6</td>
<td>A</td>
<td>Nonsynonymous</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>852682</td>
<td>Rv0758</td>
<td>pimA</td>
<td>C</td>
<td>Nonsynonymous</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>1310662</td>
<td>Rv1179c</td>
<td>iicA6</td>
<td>T</td>
<td>Nonsynonymous</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>2472593</td>
<td>Rv2208</td>
<td>cpoB</td>
<td>A</td>
<td>Synonymous</td>
<td>A</td>
<td>C</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>2706379</td>
<td>Rv2408</td>
<td>PE24</td>
<td>C</td>
<td>Synonymous</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>3176383</td>
<td>Rv2831</td>
<td>ecbA6</td>
<td>C</td>
<td>Nonsynonymous</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>3200398</td>
<td>Rv2896c</td>
<td>A</td>
<td>C</td>
<td>Nonsynonymous</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>3902781</td>
<td>Rv3137</td>
<td>Rv3137</td>
<td>A</td>
<td>Nonsynonymous</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>3975994</td>
<td>Rv3535c</td>
<td>Rv3535</td>
<td>C</td>
<td>Nonsynonymous</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>3975488</td>
<td>Rv3537</td>
<td>Rv3537</td>
<td>C</td>
<td>Synonymous</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>3997479</td>
<td>Rv3557c</td>
<td>Rv3557c</td>
<td>A</td>
<td>Nonsynonymous</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>4218300</td>
<td>Rv3773c</td>
<td>Rv3773c</td>
<td>A</td>
<td>Nonsynonymous</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>567913</td>
<td>N/A</td>
<td>N/A</td>
<td>T</td>
<td>Intergenic</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>600590</td>
<td>Rv0509</td>
<td>hemK</td>
<td>G</td>
<td>Nonsynonymous</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>809416</td>
<td>Rv0712</td>
<td>Rv0712</td>
<td>A</td>
<td>Nonsynonymous</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>1368046</td>
<td>Rv1217c</td>
<td>Rv1217c</td>
<td>C</td>
<td>Nonsynonymous</td>
<td>C</td>
<td>T</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>1837767</td>
<td>Rv1683</td>
<td>avrB</td>
<td>G</td>
<td>Synonymous</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>1917190</td>
<td>Rv1692</td>
<td>Rv1692</td>
<td>G</td>
<td>Nonsynonymous</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>1973649</td>
<td>Rv1747</td>
<td>Rv1747</td>
<td>C</td>
<td>Synonymous</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>2282000</td>
<td>N/A</td>
<td>N/A</td>
<td>T</td>
<td>Intergenic</td>
<td>T</td>
<td>T</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>2501354</td>
<td>Rv2227</td>
<td>Rv2227</td>
<td>G</td>
<td>Nonsynonymous</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>2625818</td>
<td>N/A</td>
<td>N/A</td>
<td>G</td>
<td>Intergenic</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>2761037</td>
<td>Rv2458</td>
<td>momM</td>
<td>A</td>
<td>Synonymous</td>
<td>A</td>
<td>C</td>
<td>G</td>
<td>C</td>
</tr>
<tr>
<td>2783792</td>
<td>Rv2477c</td>
<td>Rv2477c</td>
<td>G</td>
<td>Nonsynonymous</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>2970814</td>
<td>Rv2653c</td>
<td>Rv2653c</td>
<td>A</td>
<td>Synonymous</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>A</td>
</tr>
<tr>
<td>3108453</td>
<td>Rv2800</td>
<td>Rv2800</td>
<td>T</td>
<td>Synonymous</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>3374067</td>
<td>Rv3014c</td>
<td>ligA</td>
<td>G</td>
<td>Nonsynonymous</td>
<td>A</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>3430937</td>
<td>Rv3066</td>
<td>Rv3066</td>
<td>G</td>
<td>Nonsynonymous</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>3467000</td>
<td>Rv3097c</td>
<td>lipY</td>
<td>T</td>
<td>Nonsynonymous</td>
<td>T</td>
<td>T</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>3599865</td>
<td>Rv3224</td>
<td>Rv3224</td>
<td>G</td>
<td>Nonsynonymous</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>3732940</td>
<td>Rv3360</td>
<td>Rv3360</td>
<td>G</td>
<td>Nonsynonymous</td>
<td>G</td>
<td>C</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>4322291</td>
<td>Rv3785</td>
<td>Rv3785</td>
<td>C</td>
<td>Nonsynonymous</td>
<td>G</td>
<td>C</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>4297809</td>
<td>Rv3822</td>
<td>Rv3822</td>
<td>G</td>
<td>Nonsynonymous</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>4407869</td>
<td>Rv3921c</td>
<td>Rv3921c</td>
<td>G</td>
<td>Nonsynonymous</td>
<td>G</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>1558108</td>
<td>Rv1384</td>
<td>carB</td>
<td>C</td>
<td>Synonymous</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>C</td>
</tr>
<tr>
<td>3644579</td>
<td>Rv2263</td>
<td>Rv2263</td>
<td>G</td>
<td>Nonsynonymous</td>
<td>G</td>
<td>G</td>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td>921390</td>
<td>Rv0828c</td>
<td>Rv0828c</td>
<td>T</td>
<td>Nonsynonymous</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>2085436</td>
<td>Rv1835c</td>
<td>Rv1835c</td>
<td>C</td>
<td>Nonsynonymous</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
</tbody>
</table>
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 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 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 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 re-emergence 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

---

**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.
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

---

Table 1. Household and Social Contacts With Active Tuberculosis of the Same Genotype for Each Smear-Positive Case by WGS Epidemiologic Subgroup

<table>
<thead>
<tr>
<th>Subgroup by WGS and Epidemiology</th>
<th>Date of Diagnosis</th>
<th>Smear-Positive Cases</th>
<th>Smear Grade</th>
<th>Contacts With Same Genotype/Total Contacts, No. (%)&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>IIIA (n = 1)</td>
<td>May 2012</td>
<td>. . .</td>
<td>3+</td>
<td>Household Contacts: 0/0, Social Contacts: 0/18 (0)</td>
</tr>
<tr>
<td>IIIA (n = 2)</td>
<td>November 2011</td>
<td>November 2011</td>
<td>3+</td>
<td>1/4 (25)</td>
</tr>
<tr>
<td>IIIA (n = 8)</td>
<td>May 2012</td>
<td>June 2012</td>
<td>4+</td>
<td>0/0, 3/10 (30)</td>
</tr>
<tr>
<td>IIIB (n = 5)</td>
<td>December 2011</td>
<td>December 2011</td>
<td>4+</td>
<td>0/0, 4/32 (13)</td>
</tr>
<tr>
<td>IIIB (n = 13)</td>
<td>March 2012</td>
<td>May 2012</td>
<td>2+</td>
<td>2/2 (100), 3/21 (14)</td>
</tr>
<tr>
<td>IIIC (n = 20)</td>
<td>January 2012</td>
<td>January 2012</td>
<td>3+</td>
<td>3/10 (30), 12/31 (39)</td>
</tr>
<tr>
<td>Total</td>
<td>. . .</td>
<td>. . .</td>
<td>. . .</td>
<td>10/18 (56), 42/219 (19)</td>
</tr>
</tbody>
</table>

Abbreviation: WGS, whole-genome sequencing.

<sup>a</sup> Smear positive was defined as 1+ or higher, except the first subgroup comprised only 1 person, who had smear-negative disease.

<sup>b</sup> Because different sources named the same contacts, the denominators of contacts who developed active tuberculosis exceed the number of unique cases in the year.

<sup>c</sup> A 2-sample z test was used to assess difference in proportions (P < .001).
minimum coverage of 21×, and 97% of the SNPs identified had a Phred score of >100, equivalent to a 1 in 10^8 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 copypedited. 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 Palaq-sivik 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**


