

**Unravelling the Thread of Transmission: Pairing Traditional Epidemiology with Genomics
to Understand Tuberculosis Transmission Dynamics**

by

Jennifer L. Guthrie

B.Sc. (Honours), Brock University, 2000

M.Sc. (Distinction), Brock University, 2003

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Population and Public Health)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

December 2018

© Jennifer L. Guthrie, 2018

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

Unravelling the Thread of Transmission: Pairing Traditional Epidemiology with Genomics to Understand Tuberculosis Transmission Dynamics

submitted by Jennifer L. Guthrie in partial fulfillment of the requirements for

the degree of Doctor of Philosophy

in Population and Public Health

Examining Committee:

Dr. Jennifer Gardy, School of Population and Public Health

Supervisor

Dr. James Johnston, TB Services Evaluation Lead, BC Centre for Disease Control

Supervisory Committee Member

Dr. Bonnie Henry, BC Provincial Health Officer

Supervisory Committee Member

Dr. Martin Hirst, Microbiology and Immunology

University Examiner

Dr. Richard Harrigan, Experimental Medicine

University Examiner

Additional Supervisory Committee Members:

Dr. Lindsay Eltis, Microbiology & Immunology

Supervisory Committee Member

Supervisory Committee Member

Abstract

Background: In Canada, TB remains a public health concern, with the disease becoming increasingly entrenched in our most vulnerable populations. In 2012, British Columbia (BC) prepared a strategic plan with the aim of reducing incidence in the province by 50% over 10 years. A key aspect of this is preventing person-to-person spread of TB within BC—challenging, as our understanding of endemic transmission is incomplete. The objective of this dissertation is to use new advances in technology, including whole genome sequencing (WGS), to address the knowledge gaps around who, where, and how transmission occurs in BC and provide the foundation upon which new TB prevention policies and programs will be developed.

Methods: This dissertation draws on data routinely collected from persons diagnosed with culture-confirmed TB in BC (2005–2014) and linked to laboratory data from their corresponding *Mycobacterium tuberculosis* (*Mtb*) isolate(s). Collaborations with Yukon and Ontario provided additional data. 24-locus MIRU-VNTR genotype results were available for each *Mtb* isolate across all three study populations. WGS of genotypically clustered BC isolates and all Yukon isolates was carried out and the data analyzed using a bioinformatics pipeline developed at Oxford University. Descriptive and inferential statistics were used to examine clinical and demographic characteristics of persons with *Mtb* isolates belonging to a cluster according to genotyping and WGS methods.

Results: Universal genotyping of all *Mtb* isolates collected in BC over a ten-year period revealed that 57.6% of the study population had a genotypically unique *Mtb* isolate. Sixteen large genotype clusters were identified, nine in predominately Canadian-born (CB) persons. Application of WGS indicated the large genotypic clusters comprised of mainly non-Canadian-born (nCB) persons did not represent recent, endemic transmission within BC, and WGS additionally refined many of the CB clusters to smaller sub-clusters. The WGS clustered proportion was 25.8%.

Conclusions: Approximately one in four of BC's TB cases occur in CB persons and are largely the result of local transmission. WGS represents a new and important tool for understanding the spread of TB within a population, and using this technology paired with routinely collected case-level data provided significant insights to transmission in BC.

Lay Summary

With an estimated 1.7 million deaths annually, tuberculosis (TB) is the number one infectious disease killer worldwide. In Canada, TB remains a public health concern, with the disease becoming increasingly entrenched in our most vulnerable populations. Recent advances now allow the full DNA sequence to be read from TB bacteria isolated from persons diagnosed with the disease. Analysis and comparison of tuberculosis DNA sequences from persons with TB in British Columbia (BC) over a ten-year period was used to detect small changes in the DNA as TB is spread (transmitted) from person-to-person. By combining this information with case-level data important insights to tuberculosis transmission were gained, including risk factors related to TB spread and a more accurate estimate of disease resulting from transmission within BC. Understanding person-to-person spread of TB is crucial to improving awareness, education, early detection, and ultimately preventing the spread of tuberculosis in the most at-risk communities.

Preface

All the works presented henceforth were conducted at the University of British Columbia (UBC), Vancouver, and at the British Columbia Centre for Disease Control (BCCDC), Vancouver, BC, Canada, and conceived, undertaken, and written by the candidate, Jennifer L. Guthrie (JLG_u). Data for the research projects presented within have been obtained from several distinct sources: (1) BCCDC Public Health Laboratory (2) BC's tuberculosis registry—Integrated Public Health Information System (iPHIS) (3) Yukon's tuberculosis registry (iPHIS) (4) Immigration, Refugees and Citizenship Canada Permanent Residents database records, linked to data from Population Data BC health administrative databases, and (5) Public Health Ontario Laboratories. Data linkage for each study was performed under the BCCDC's public health mandate with approval of the Privacy and Access Committee. The UBC Clinical Research Ethics Board provided ethical approvals (#H12-00910, #H16-00265). An additional ethics approval (#2016-058.0) was obtained for the collaboration with Public Health Ontario (**Chapter 4**).

The co-authors of the manuscripts, including Dr. Jennifer L. Gardy (JLG_a), Dr. James Johnston (JJ), Dr. Victoria J. Cook (VJC), Dr. Patrick Tang (PT), Dr. Linda Hoang (LH), Dr. Kevin Elwood (KE), Dr. David Roth (DR), Dr. Bonnie Henry (BHe), Dr. Andy Delli Pizzi (ADP), Dr. Sarah Cherian (SC), Dr. Lisa Ronald (LR), Dr. Mabel Rodrigues (MR), Clare Kong (CK), Danielle Jorgensen (DJ), Maichael Thejoe (MT), Dr. Frances B. Jamieson (FBJ), Alex Marchand Austin (AMA), Karen Lam (KL), Daria Pyskir (DP), Kirby Cronin (KC), Dr. Brendan Hanley (BHa), Lori Strudwick (LS), Beth Roberts (BR), Meadow Allen (MA), Jan McFadzen (JM), Dr. Timothy Walker (TW) and Dona Foster (DF) made contributions only as is commensurate with collegial or co-investigator duties. Relative contributions of the author, collaborators, and co-authors are described in detail below.

Chapter 1, Sections 1.1 – 1.5 and 0, Sections 9.1 – 9.3, 9.5 – 9.6 are the original, unpublished intellectual products of the candidate. With guidance and input from my supervisor (JLG_a), the literature was searched and reviewed, and the findings of these chapters synthesized.

A version of **Chapter 1, Section 1.6** has been published. **Guthrie JL**, and Gardy, JL. (2017). A brief primer on genomic epidemiology: lessons learned from *Mycobacterium tuberculosis*. *Annals of the New York Academy of Sciences*. 1388:59–77. [dx.doi.org/10.1111/nyas.13273](https://doi.org/10.1111/nyas.13273). JLGu in collaboration with JLGa developed the manuscript concept, conducted a literature review, drafted and revised the manuscript.

A version of **Chapter 1, Section 1.7** has been published. **Guthrie JL**, and Gardy JL. (2015). Accelerating tuberculosis elimination in low-incidence settings: the role of genomics. *European Respiratory Journal*. 46:1840–1841. [dx.doi.org/10.1183/13993003.00788-2015](https://doi.org/10.1183/13993003.00788-2015). JLGu in collaboration with JLGa developed the manuscript concept, conducted a literature review, drafted and revised the manuscript.

A version of **Chapter 2** has been published. **Guthrie JL**, Kong C, Roth D, Jorgensen D, Rodrigues M, Tang P, Thejoe M, Elwood K, Cook VJ, Johnston J, Gardy JL. (2018). Universal Genotyping for Tuberculosis Prevention Programs: A Five-Year Comparison with On-Request Genotyping. *Journal of Clinical Microbiology*. 56:e01778-17. JLGu developed the study, linked, cleaned and analyzed the data, drafted and revised the manuscript. JLGa supervised the study. CK carried out the laboratory work under the supervision of MR and PT. DR extracted the provincial case data. CK, DR, DJ, MR, PT, MT, KE, VJC, JJ, and JLGa provided input to the draft and contributed to the revision of the final manuscript.

A version of **Chapter 3** has been published. **Guthrie, JL**, Kong, C, Roth, D, Jorgensen, D, Rodrigues, M, Hoang, L, Tang, P, Cook, V, Johnston, J, and Gardy, JL. (2017). Molecular Epidemiology of Tuberculosis in British Columbia, Canada: A 10-Year Retrospective Study. *Clinical Infectious Diseases*. 66:849–856. JLGu conceived of and designed the study, linked, cleaned and analyzed the data, drafted and revised the manuscript. JLGa supervised the study. CK carried out the laboratory work under the supervision of MR and PT. DR extracted the provincial case data. CK, DR, DJ, MR, LH, PT, VJC, JJ, and JLGa provided input to the draft and contributed to the revision of the final manuscript.

A version of **Chapter 4** has been submitted for peer review. **Guthrie JL**, Marchand-Austin A, Cronin K, Lam K, Pyskir D, Kong C, Jorgensen D, Rodrigues M, Roth D, Tang P, Cook VJ, Johnston J, Jamieson FB, Gardy JL. (2018). Universal Genotyping Reveals Province-Level Differences in the Molecular Epidemiology of Tuberculosis. *Manuscript submitted*. This study is based on a collaboration with Public Health Ontario (PHO) and lead by JLGa. A collaboration agreement between BCCDC and PHO was approved on May 31, 2018 (#RRB-17-025). JLGa conceived of and designed the study, linked, cleaned and analyzed the data, drafted and revised the manuscript. JLGa supervised the study. CK carried out the BC laboratory work under the supervision of MR and PT. KL and DP carried out the ON laboratory work under the supervision of FJ. KC extracted the ON provincial case data under the supervision of AMA. AMA, KC, KL, DP, CK, DJ, MR, DR, PT, VJC, JJ, FBJ and JLGa provided input to the draft and contributed to the revision of the final manuscript.

A version of **Chapter 5** has been published. **Guthrie JL**, Delli Pizzi A, Roth D, Kong C, Jorgensen D, Rodrigues M, Tang P, Cook VJ, Johnston J, and Gardy JL. (2018). Genotyping and Whole Genome Sequencing to Identify Tuberculosis Transmission to Pediatric Patients in British Columbia, Canada, 2005–2014. *The Journal of Infectious Diseases*. 218:1155–1163. JLGa conceived of and designed the study, linked, cleaned and analyzed the data, drafted and revised the manuscript. ADP reviewed pediatric case notes and provided a summary of contact investigations. JLGa supervised the study. CK carried out the laboratory work under the supervision of MR and PT. DR extracted the provincial case data. ADP, DR, CK, DJ, MR, PT, VJC, JJ and JLGa provided input to the draft and contributed to the revision of the final manuscript.

A version of **Chapter 6** has been submitted for peer review. **Guthrie JL**, Strudwick L, Roberts B, Allen M, McFadzen J, Roth D, Jorgensen D, Rodrigues M, Tang P, Hanley B, Johnston J, Cook VJ, Gardy JL. (2018). Whole Genome Sequencing for Improved Understanding of *Mycobacterium tuberculosis* Transmission in a Remote Circumpolar Region. *Submitted*. **Chapter 6** is based on a collaboration with Yukon Communicable Disease Control (YCDC) and lead by JLGa under the supervision of JLGa, VJC (BCCDC Physician Consultant to Yukon) and

BH (Yukon Chief Medical Officer of Health). YCDC contracts BCCDC for laboratory work and TB services support; however, to formalize the collaboration for this research project an information sharing agreement between BCCDC and YCDC was prepared and approved on October 5, 2017 (#2017-008). JLGu conceived of and designed the study, linked, cleaned and analyzed the data, drafted and revised the manuscript. CK carried out the laboratory work under the supervision of MR and PT. LR and DR extracted the YT provincial case data. LS, BR, MA, and JM reviewed YT cases and summarized contact investigation notes. LS, BR, MA, JM, DR, DJ, MR, PT, BH, JJ, VJC and JLGa provided input to the draft and contributed to the revision of the final manuscript.

A version of **Chapter 7** is currently being finalized for submission. **Guthrie JL**, Strudwick L, Roberts B, Allen M, McFadzen J, Roth D, Jorgensen D, Rodrigues M, Tang P, Hanley B, Johnston J, Cook VJ, Gardy JL. (2018). Comparison of Traditional Field Epidemiology and Whole Genome Sequencing to Understand Tuberculosis Transmission in a Remote. *Manuscript in preparation*. **Chapter 7** is based on a collaboration with Yukon Communicable Disease Control (YCDC) and lead by JLGu under the supervision of JLGa, VJC (BCCDC Physician Consultant to Yukon) and BH (Yukon Chief Medical Officer of Health). YCDC contracts BCCDC for laboratory work and TB services support; however, to formalize the collaboration for this research project an information sharing agreement between BCCDC and YCDC was prepared and approved on October 5, 2017 (#2017-008). JLGu conceived of and designed the study, linked, cleaned and analyzed the data, drafted and revised the manuscript. JLGu developed the online and in-person surveys with input from JLGa and VJC. LS, BR, MA, and JM participated in the online and in-person surveys. JLGu and JLGa lead in-person case discussions, qualitative survey and data collection. CK carried out the laboratory work under the supervision of MR and PT. LR and DR extracted the YT provincial case data. LS, BR, MA, and JM reviewed YT cases and summarized contact investigation notes. LS, BR, MA, JM, DR, DJ, MR, PT, BH, JJ, VJC and JLGa provided input to the draft and contributed to the revision of the final manuscript.

A version of **Chapter 8** is currently being prepared for submission. **Guthrie JL**, Cherian S, Kong C, Roth D, Jorgensen D, Rodrigues M, Walker T, Foster D, Henry B, Cook VJ, Johnston J, Tang P, Gardy JL. (2018). Whole Genome Sequencing as a Tool to Understand and Quantify Active Tuberculosis Arising from Local Transmission. *Manuscript in preparation*. JLGu conceived of the study with JLGa, and designed the analysis protocols in consultation with JLGa, PT, VJC, and JJ. SC reviewed case notes and extracted case data. JLGu linked, cleaned and analyzed the data, drafted and revised the manuscript. JLGa supervised the study. CK carried out the laboratory work under the supervision of MR and PT. DR extracted the provincial case data. JLGu and JLGa reviewed case notes where necessary. TM and DF provided support and coordination of genomic analysis using the Oxford University bioinformatics pipeline. SC, CK, DR, DJ, MR, TW, DF, BH, VJC, JJ, PT and JLGa are providing input to the draft and will contribute to the revision of the final manuscript.

A version of **Chapter 9, Section 9.3.1** is currently undergoing peer review. **Guthrie JL**, Ronald L, Cook VJ, Johnston J, Gardy, JL. (2018). Homogeneous Group or a Multicultural Mosaic? The Challenge with Reporting Birth Outside Canada as a Tuberculosis Risk. *Submitted*. JLGu in collaboration with JLGa developed the manuscript concept with input from LR, VJC and JJ. JLGu analyzed the genotyped data set. LR contributed results from the CIC/PopData BC data set. JLGu drafted and revised the manuscript. LR, VJC, JJ, and JLGa provided input to the draft and contributed to the revision of the final manuscript.

Table of Contents

Abstract.....	iii
Lay Summary	iv
Preface.....	v
Table of Contents	x
List of Tables	xviii
List of Figures.....	xx
List of Abbreviations	xxiii
Glossary	xxvi
Acknowledgements	xxvii
Chapter 1: Introduction	1
1.1 Background and Rationale.....	1
1.2 Research Objectives.....	3
1.3 Tuberculosis.....	5
1.3.1 Etiology.....	5
1.3.2 Diagnosis and treatment.....	6
1.3.3 <i>Mycobacterium tuberculosis</i> lineages.....	8
1.3.4 Transmission.....	9
1.3.5 Monitoring, surveillance and investigation.....	10
1.3.6 Epidemiology of tuberculosis in Canada	11
1.4 Genotyping of <i>Mycobacterium tuberculosis</i>	13
1.4.1 RFLP.....	13

1.4.2	MIRU-VNTR.....	13
1.4.3	Spoligotyping.....	14
1.5	Molecular Epidemiology and Public Health.....	15
1.6	A Brief Primer on Genomic Epidemiology: Lessons Learned from <i>Mycobacterium tuberculosis</i>	16
1.6.1	The <i>Mycobacterium tuberculosis</i> genome.....	16
1.6.2	Leveraging genomics for epidemiology.....	17
1.6.3	Step 1: look before you leap.....	23
1.6.4	Step 2: from sample to sequence.....	26
1.6.5	Step 3: bases to bytes.....	29
1.6.6	Step 4: rapid resistance prediction.....	35
1.6.7	Step 5: making the links.....	37
1.6.8	Concluding thoughts.....	40
1.7	Accelerating TB Elimination in Low-incidence Settings: the Role of Genomics.....	41
Chapter 2: Universal Genotyping for Tuberculosis Prevention Programs: A Five-Year Comparison with On-Request Genotyping.....		44
2.1	Background.....	44
2.2	Materials and Methods.....	45
2.2.1	On-request genotyping data.....	45
2.2.2	Universal genotyping data.....	45
2.2.3	Statistical analysis.....	46
2.3	Results.....	47

2.3.1	The genotype request proportion was smaller than the genotypic clustering proportion.....	47
2.3.2	Requests reflected suspected community transmission and known risk factors.....	49
2.3.3	Universal genotyping improves cluster identification	51
2.3.4	Growing clusters were variably identified by on-request genotyping.....	55
2.4	Discussion.....	57
Chapter 3: Molecular Epidemiology of Tuberculosis in British Columbia, Canada—A 10-Year Retrospective Study.....		60
3.1	Background.....	60
3.2	Materials and Methods.....	61
3.2.1	Study setting and design	61
3.2.2	Case data	61
3.2.3	Laboratory analysis.....	62
3.2.4	Statistical analysis.....	63
3.3	Results.....	63
3.3.1	Lineage analysis.....	68
3.3.2	MIRU-VNTR identifies discrete subgroups amongst BC’s TB cases.....	72
3.3.3	MIRU-VNTR identifies drivers of large transmission clusters	75
3.4	Discussion.....	76
Chapter 4: Universal Genotyping Reveals Province-Level Differences in the Molecular Epidemiology of Tuberculosis.....		81
4.1	Background.....	81
4.2	Methods.....	82

4.2.1	Study setting and design	82
4.2.2	Diagnosis and case information	83
4.2.3	Genotyping by 24-locus MIRU-VNTR	83
4.2.4	Statistical analysis.....	84
4.3	Results.....	84
4.3.1	Descriptive epidemiology	84
4.3.2	TB isolates in BC are more likely to be clustered by MIRU-VNTR.....	86
4.3.3	Interprovincial clustering occurs frequently between Ontario and BC	87
4.4	Discussion.....	96
 Chapter 5: Genotyping and Whole-Genome Sequencing to Identify Tuberculosis		
Transmission to Pediatric Patients in British Columbia.....99		
5.1	Background.....	99
5.2	Methods.....	100
5.2.1	Study setting and design	100
5.2.2	Case data	101
5.2.3	Laboratory methods	101
5.2.4	Whole genome sequencing analysis	102
5.2.5	Transmission classification.....	103
5.2.6	Statistical analysis.....	103
5.3	Results.....	104
5.3.1	Demographics, clinical presentation, and epidemiology	104
5.3.2	Molecular and genomic epidemiology investigation of putative sources.....	107
5.3.3	Identification of infections acquired out of province.....	108

5.3.4	Identification of locally acquired infections	110
5.3.5	Household transmission of multidrug resistant tuberculosis	113
5.4	Discussion.....	113
Chapter 6: Whole Genome Sequencing for Improved Understanding of <i>Mycobacterium tuberculosis</i> Transmission in a Remote Circumpolar Region.....117		
6.1	Background.....	117
6.2	Methods.....	118
6.2.1	Study setting and design	118
6.2.2	Case-level information.....	119
6.2.3	Laboratory methods	120
6.2.4	WGS analysis.....	120
6.2.5	Statistical Analysis.....	121
6.3	Results.....	121
6.3.1	MIRU-VNTR and WGS provide different estimates of clustering	121
6.3.2	Genomically related cases across jurisdictions are similar clinically.....	125
6.3.3	Transmission reconstruction	127
6.4	Discussion.....	130
Chapter 7: Comparison of Traditional Field Epidemiology and Whole Genome Sequencing to Understand Tuberculosis Transmission in a Remote Setting133		
7.1	Background.....	133
7.2	Methods.....	135
7.2.1	Study setting and design	135
7.2.2	Bacterial culture, genotyping and whole genome sequencing.....	135

7.2.3	Source identification by field and molecular epidemiology	136
7.2.4	Source identification by genomic epidemiology	137
7.2.5	Source identification consensus	137
7.2.6	Qualitative assessment	138
7.2.7	Statistical methods	138
7.3	Results.....	139
7.3.1	Good agreement around clusters and location of TB exposure between methods .	139
7.3.2	Low genomic variability within clusters limited of an exact source	143
7.3.3	Confidence in correct source identification varied between teams	144
7.3.4	Preference for genomics over genotyping	146
7.4	Discussion.....	149

Chapter 8: Whole Genome Sequencing as a Tool to Understand and Quantify Active

Tuberculosis Arising from Local Transmission.....152

8.1	Background.....	152
8.2	Methods.....	153
8.2.1	Study population	153
8.2.2	Case data	153
8.2.3	Laboratory analysis.....	153
8.2.4	WGS analysis and genomic clustering	154
8.2.5	Transmission across population groups	155
8.2.6	Tuberculosis reoccurrences.....	155
8.2.7	Characterization of large clusters and transmission reconstruction.....	156
8.2.8	Statistical analysis.....	156

8.3	Results.....	157
8.3.1	Whole genome sequencing reduces the local transmission estimate.....	161
8.3.2	Transmitted <i>Mtb</i> isolates belong largely to the Euro-American lineage	164
8.3.3	Risk factors for local transmission.....	164
8.3.4	Transmission occurs in both directions between Canadian-born and non-Canadian-born persons	168
8.3.5	TB relapse vs reinfection	171
8.3.6	Characterization of large genomic clusters.....	173
8.4	Discussion.....	178
Chapter 9: Conclusion.....		182
9.1	Summary of findings.....	182
9.2	Unique contributions, implications and impact	185
9.3	Strengths and limitations.....	187
9.3.1	Homogeneous Group or a Multicultural Mosaic? The Challenge with Reporting Birth Outside Canada as a Tuberculosis Risk.....	190
9.4	Knowledge translation	197
9.5	Future research and recommendations.....	198
9.5.1	Prospective provincial MIRU-VNTR genotyping.....	198
9.5.2	Standardization of WGS bioinformatics pipelines	198
9.5.3	WGS as a tool for TB prevention	199
9.6	Final Conclusions.....	200
References.....		201
Appendices.....		238

Appendix A: Online Pre- and Post-Meeting Survey Questions.....	238
Appendix B: Presentations.....	242
B.1 Oral presentations	242
B.2 Poster Presentations	243
B.3 Media	243
Appendix C: Genotyping Summary Report.....	244

List of Tables

Table 1-1. Genomic epidemiology TB studies examining transmission between individuals.....	22
Table 1-2. Sequencing platforms currently deployed for genomic epidemiology studies	28
Table 2-1. Genotype request reasons	48
Table 2-2. Study sample characteristics.....	50
Table 2-3. Characteristics of MIRU-VNTR clusters	51
Table 2-4. Logistic regression.....	53
Table 2-5. Logistic regression with a restricted dataset.....	54
Table 2-6. Request status, risk factor, and clustering	55
Table 2-7. Genotype cluster characteristics	56
Table 3-1. Study population.....	65
Table 3-2. Multi-drug resistant isolates	67
Table 3-3. Lineage by anatomical disease site.....	70
Table 3-4. Large cluster characteristics	71
Table 3-5. Genotype cluster sizes	73
Table 3-6. Risk factors for genotypic clustering.....	74
Table 3-7. Risk factors associated with cluster size	76
Table 4-1. Study population.....	85
Table 4-2. Genotype cluster sizes	87
Table 4-3. Multivariable logistic regression	90
Table 4-4. Multivariable logistic regression according to size.....	91
Table 4-5. Large genotypic clusters.....	94

Table 5-1. Demographic and clinical characteristics of culture-positive pediatric TB cases	106
Table 5-2. Factors associated with locally acquired pediatric tuberculosis.....	112
Table 6-1. Characteristics of Yukon study population	126
Table 7-1. Location of TB Acquisition.....	142
Table 7-2. High level concordance between methods	142
Table 7-3. Concordance between methods at a case-level	143
Table 7-4. Accuracy of source case identification.....	145
Table 8-1. Characteristics of the study sample	159
Table 8-2. Genomic clustering logistic regression	166
Table 8-3. Logistic regression for risk factors for genomic clustering—various thresholds.....	167
Table 8-4. Large genomic clusters.....	174

List of Figures

Figure 1-1. Tuberculosis incidence (per 100,000 population) in Canada, 1924–2014.....	11
Figure 1-2. Tuberculosis incidence (per 100,000 population) by Canadian province/territory, 2014.....	12
Figure 1-3. Molecular epidemiology methods used in tuberculosis surveillance.....	18
Figure 1-4. The basic principle of genomic epidemiology.....	20
Figure 1-5. Identifying transmission-informative variation.....	34
Figure 2-1. Study sample request status	47
Figure 2-2. Quarterly genotype requests.....	48
Figure 2-3. Proportion of each cluster requested by cluster growth over time.....	52
Figure 2-4. Cluster growth by genotype request status.....	57
Figure 3-1. Molecular epidemiology study inclusion/exclusion criteria	62
Figure 3-2. Population structure of <i>Mycobacterium tuberculosis</i> genotypes in BC.....	69
Figure 3-3. <i>Mycobacterium tuberculosis</i> lineage by continent of birth.....	70
Figure 4-1. Genotypes shared between provinces	88
Figure 4-2. Proportion of genotypic clustering.....	88
Figure 4-3. Interprovincial genotype matches	92
Figure 4-4. Single contributors to clusters.....	92
Figure 4-5. Population structure of <i>Mycobacterium tuberculosis</i> isolates shared between BC and Ontario	95
Figure 5-1. Pediatric TB study inclusion/exclusion criteria	104
Figure 5-2. Pediatric age distribution by birthplace.....	105

Figure 5-3. Number of contacts	107
Figure 5-4. Pediatric tuberculosis investigation summary.....	109
Figure 5-5. Pediatric analysis phylogenetic tree.....	111
Figure 6-1. Study sample	119
Figure 6-2. Population structure of <i>Mycobacterium tuberculosis</i> in Yukon Territory	122
Figure 6-3. Yukon Territory <i>Mycobacterium tuberculosis</i> isolates in the context of related BC isolates.....	124
Figure 6-4. Yukon WGS transmission reconstructions	128
Figure 6-5. Transmission clusters—SNV alignments	130
Figure 7-1. MIRU-VNTR cluster summary example.....	136
Figure 7-2. Yukon cases over time	139
Figure 7-3. Whole genome sequencing-based population structure of Yukon Territory <i>Mycobacterium tuberculosis</i> isolates.....	141
Figure 7-4. Certainty assigned to identified sources.....	144
Figure 7-5. Frequency of certainty categories assigned for each source identified, divided by cluster.....	145
Figure 8-1. Study sample inclusion/exclusion criteria.....	158
Figure 8-2. Maximum-likelihood phylogenetic tree of study isolates	162
Figure 8-3. Genomic cluster sizes.....	163
Figure 8-4. Pairwise SNV distances between study isolates	163
Figure 8-5. Genomic cluster sizes and birthplace composition	169
Figure 8-6. Mixed cluster transmission	170
Figure 8-7. Recurrent tuberculosis characteristics.....	171

Figure 8-8. Timeline of case diagnosis	176
Figure 8-9. Characterization of WClust-2	177
Figure 9-1. Trends in active tuberculosis diagnoses in British Columbia (BC), Canada	192
Figure 9-2. Tuberculosis incidence rates in British Columbia (BC) for persons born outside of Canada.....	193
Figure 9-3. Number of tuberculosis cases for each large (≥ 10 persons) genotypic cluster	194

List of Abbreviations

AFB	acid-fast bacilli
AIC	Akaike’s Information Criterion
aOR	adjusted odds ratio
BC	British Columbia
BCG	Bacillus Calmette–Guérin
BCCDC	British Columbia Centre for Disease Control
BCPHL	British Columbia Centre for Disease Control Public Health Laboratory
bp	base pair
BED	browser extensible data
BAM	Binary Alignment Map
CAN-Marg	Canadian Marginalization Index
CBP	Canadian-born parents
CDC	Centers for Disease Control and Prevention
CI	confidence interval
CI	contact investigations
CRyPTIC	Comprehensive Resistance Prediction for Tuberculosis: an International Consortium
DA	dissemination area
DNA	deoxyribonucleic acid
DR	direct repeat
DST	drug susceptibility testing
DTES	Downtown Eastside
GVR	Greater Vancouver Region
HIV	human immunodeficiency virus
INH	isoniazid
IGRA	Interferon gamma release assay
iPHIS	Integrated Public Health Information System
IQR	interquartile range

LJ	Lowenstein-Jensen
LIMS	laboratory information management systems
LTBI	latent tuberculosis infection
MCAR	missing completely at random
MClustID	MIRU-VNTR cluster identifier
MDR	multi-drug resistance
MGEs	mobile genetic elements
MGIT	Mycobacteria growth indicator tube
MIRU-VNTR	Mycobacterial Interspersed Repetitive Units-Variable Number Tandem Repeats
MST	minimum-spanning tree
<i>Mtb</i>	<i>Mycobacterium tuberculosis</i>
nCB	non-Canadian-born
NCBI	National Center for Biotechnology Information
nCBP	non-Canadian-born parents
NRTB	exclusively non-respiratory tuberculosis
ON	Ontario
OR	odds ratio
PCR	polymerase chain reaction
PHAC	Public Health Agency of Canada
PHE	Public Health England
PHO	Public Health Ontario
PZA	pyrazinamide
RFLP	restriction fragment length polymorphism
RIF	rifampin
RTB	exclusively respiratory tuberculosis
SD	standard deviation
SNV(s)	single nucleotide variant(s)
TB	tuberculosis
TST	tuberculin skin test
UBC	University of British Columbia

UK	United Kingdom
USA	United States of America
VCF	variant call format
WClustID	whole genome sequencing cluster identifier
WGS	whole genome sequencing
YCDC	Yukon Communicable Disease Control
YT	Yukon Territory

Glossary

Acid-fast bacilli (AFB) smear microscopy	Microscopic examination of a clinical specimen (e.g. sputum) prepared using a fluorochrome stain and smeared onto a glass slide to detect acid-fast bacilli.
Bacille Calmette-Guérin (BCG) vaccine	BCG is a live attenuated vaccine named after the doctors who developed it. Derived from a strain of <i>Mycobacterium bovis</i> , BCG is mainly used as a vaccination against tuberculosis, although there are widely varying results in BCG efficacy studies.
Downtown Eastside (DTES)	Neighbourhood in Vancouver with extreme poverty, homelessness, illicit drug use, mental illness and sex work.
Endemic	Disease or bacterial strain that is regularly found in an area or very common among a particular group.
Induration	A palpable, raised, hardened area.
Multidrug-resistant tuberculosis (MDR-TB)	Tuberculosis resistant to isoniazid and rifampin with or without resistance to other anti-tuberculosis drugs.
Reactivation	The development of active disease after a period of latent tuberculosis infection.
Reinfection	Individual previously treated (cure or completed) for active TB, in whom active tuberculosis is detected ≥ 6 months following previous treatment completion, and the new isolate is confirmed to have a difference genotype from the original organism.
Relapse	Individual previously treated (cure or completed) for active TB, in whom active tuberculosis is detected ≥ 6 months following previous treatment completion, and the new isolate is confirmed to have the same genotype as the original organism.
Reoccurrence	Individual previously treated (cure or completed) for active TB, in whom active tuberculosis is detected ≥ 6 months following previous treatment completion.
Super-spreader	An individual who transmits TB to a greater number of secondary cases than the average infected person

Acknowledgements

I would first like to thank my supervisor, Dr. Jennifer Gardy, for the patient guidance, encouragement and advice given for all things thesis and not. I have been extremely lucky to have a supervisor who cares so much—not just about the work—but also the student. Jenn fosters an open, ambitious and collaborative research culture that along with her eternal optimism and ability to deal with any obstacle, is an inspiration to my future endeavors. I am forever grateful to her for providing me with the opportunity to make my PhD goal come true.

I would also like to acknowledge the support, and guidance of my supervisory committee, Drs. James Johnston, Lindsay Eltis and Bonnie Henry. I am very much appreciative of their time, insightful comments and questions that most definitely improved my research. A big thank you to the BC Centre for Disease Control TB Services and Public Health Laboratory personnel for all their efforts, without which there would have been no thesis. In particular, Dr. Patrick Tang for his work to make this very large laboratory intensive study happen, Clare Kong who was so key to the laboratory efforts, and to Dr. David Roth for his assistance with the provincial data extraction. Also, a huge thank you to Dr. Victoria Cook for her support and big picture vision, and ever amazing discussion points. Additionally, I would like to say thank you to my collaborators at Yukon Communicable Disease Control and Public Health Ontario.

Funding was generously provided by the Canadian Institutes for Health Research through a Banting and Best Canada Graduate Scholarship, Killam Doctoral Scholarship, and the University of British Columbia Doctoral Fellowship. I also thank the BCCDC Foundation for Population and Public Health for providing project support.

I would like to express my utmost gratitude for the support and patience of my friends and family throughout my studies. This journey would not have been possible without them. Especially, my brother Dennis, Sandy, Olga, Angela, Sujay, Uncle Chuck, Aunt Gerry and the rest of the crazy but lovable Guthrie clan. My wonderful little nephews whose smiles always brighten my day, D., Hunter, Joshua and James. Also, my friends who are like family, Erynne, Lesley, Rian, Beth, Rema, Alex, David A., Christine, Ace and all the way in Wales—Charlotte and Andrew. Lastly, I would like to acknowledge my parents, my brother David, and grandma C., who always believed in my ability to persevere no matter the challenges. Although gone their belief in me remains—I know they would be proud of my accomplishments.

Chapter 1: Introduction

An individual with active respiratory tuberculosis disease, left untreated, has been estimated to infect 2–12 people per year.^{1–3} Investigating the contacts of a person with TB is a fundamental cornerstone of TB prevention and care programs, and is the key to identifying infected individuals with active disease requiring treatment, or latent TB infection that may merit prophylaxis. While it is recognized that tuberculosis transmission is driven by the interaction between host, pathogen, and environmental factors,⁴ our understanding of TB transmission in low-TB incidence settings like British Columbia is incomplete. This dissertation pairs new molecular and genomic technologies with traditional epidemiological contact investigation data to improve our knowledge and understanding of tuberculosis transmission in a low-incidence setting. An overview of the relevant history, etiology and epidemiology of tuberculosis are provided in this introductory chapter.

1.1 Background and Rationale

Tuberculosis (TB), a bacterial infection caused by *Mycobacterium tuberculosis* (*Mtb*), is a major cause of disease and death in much of the world. More than 20 years after the World Health Organization declared tuberculosis a global emergency, TB remains entrenched in the world's population. TB disease affects nine million people per year and was estimated to have caused 1.7 million deaths in 2016.⁵ While the highest burden of disease is seen in developing countries, TB remains a major public health issue in Canada. After decades of decline in TB incidence, progress towards elimination has stalled. Those born outside Canada, Indigenous persons, as well as incarcerated and under-housed individuals are at the greatest risk of presenting with active TB disease.⁶ Indeed, the TB incidence rate in certain sectors of the Canadian population can exceed that in developing countries.⁷

British Columbia has one of the highest provincial TB rates in Canada, with 6.3 cases per 100,000 population (293 cases) in 2014.⁸ Of these, persons born outside Canada accounted for

approximately 80% of BC's tuberculosis cases,⁹ many of which likely result from reactivation of latent TB infections (LTBI) acquired in the individual's country of birth. Canadian-born individuals accounted for ~20% of BC's cases, and with few exceptions these are the result of local acquisition, also referred to as endemic transmission. Nationally, endemic transmission accounts for anywhere from 10–100% of new TB cases in a province or territory.⁷ Even within a province/territory rates can vary. For example, Vancouver's Downtown Eastside (DTES), an impoverished, high-density, urban neighbourhood with a high prevalence of illicit drug use, has a very high TB incidence rate, at ~40 cases per 100,000 people,¹⁰ due almost entirely to endemic transmission.

Significant public health resources are required for TB treatment, follow-up, and contact tracing. Persons with active TB disease are often hospitalized in airborne infection isolation rooms for weeks to months and are subjected to multi-drug therapy for six months or longer with antibiotics that may have harmful side effects. Directly observed therapy may be required to facilitate individuals completing their full course of treatment, as failure to take anti-tuberculosis drugs can lead to longer illness and antibiotic-resistant TB, at which point a cure becomes significantly more challenging, if at all possible. Treating each uncomplicated active TB case is estimated to cost Canada \$47,290;¹¹ complex cases involving drug resistance can cost significantly more, and the disease exacts a physical and emotional toll on affected individuals, their contacts, and their caregivers.

In 2006, Canada set the goal of reducing the national incidence of reported TB to 3.6 cases per 100,000 population or less by 2015.¹² Although rates decreased over subsequent years, a reported incidence rate of 4.6 per 100,000 in 2015 and an increase to 4.8 the following year indicates that we failed to achieve this goal, and that we must adopt innovative new strategies to manage this disease.¹³ TB prevention and care is a shared responsibility between various levels of government, including individual provincial/territorial governments, and as such, BC developed a Provincial TB Strategy in 2012, aiming to reduce TB incidence by 50% over 10 years.¹⁴ Achieving this ambitious goal will take considerable effort, requiring innovative public health interventions specifically tailored to both prevent person-to-person transmission of tuberculosis

within BC, as well as prevent the development of active disease in individuals with LTBI by improving screening and therapy. The latter is a comparatively straightforward proposition; however, reducing active transmission within Canada is more complex, as our understanding of endemic transmission is incomplete. It is unknown precisely how many cases result from transmission of active disease and, because TB disease may not occur until years after exposure, individuals are often unable to identify the true source of their infection. This information is key to the design and delivery of effective evidence-based interventions and to prevent the continuing spread of TB. This dissertation relies on a novel approach utilizing the new technique of genomic epidemiology to fully describe the when, where, and how of endemic transmission within and between the varied regional, socioeconomic, and cultural settings of BC. The findings of this research will directly impact public health policy and practice within BC and potentially across Canada.

1.2 Research Objectives

The overarching aim of this dissertation is to describe the molecular and genomic epidemiology of TB in British Columbia and increase our understanding of the patterns underlying the person-to-person spread of TB in the province. “Molecular epidemiology” refers to the use of a rapid, low-resolution genotyping method to identify clusters of cases that might represent endemic transmission, one that is routinely used in BC’s Public Health Laboratory. “Genomic epidemiology” refers to the use of higher-resolution whole genome sequencing to more accurately identify clusters representing true endemic transmission, including inferring specific person-to-person transmission events where possible.

Specific aims and objectives include:

1. To identify gaps in the literature with regards to the use of genomics for enhanced TB care and treatment and contribute expertise in this area.
 - a. Review the literature and provide background on the use of genomic epidemiology in TB. (Manuscript 1)

- b. Communicate the role of genomics in accelerating TB elimination in low-incidence settings. (Manuscript 2)
2. To describe the molecular epidemiology of TB in British Columbia.
 - a. Assess the value of universal genotyping for the identification of clustered TB cases potentially representing local transmission. (Manuscript 3)
 - b. Describe the molecular epidemiology of TB in British Columbia using 24-locus MIRU-VNTR genotyping linked to key clinical and demographic data. (Manuscript 4)
 - c. Compare province-level TB molecular epidemiology between BC and Ontario—a similarly large, immigrant-receiving province—to identify interprovincial genotype clusters and improve our understanding of BC’s molecular epidemiology in the larger Canadian context. (Manuscript 5)
3. To calibrate molecular tools using small, well-defined populations for a more refined understanding of transmission.
 - a. Using pediatric TB cases, which often have well-defined TB exposures and extensive contact investigation data, examine person-to-person TB transmission using genotyping and WGS data. (Manuscript 6)
 - b. Combine case-level epidemiological data with genotyping and WGS analysis to identify chains of transmission within and across Yukon/BC borders. (Manuscript 7)
 - c. Using the small, well-defined cohort of Yukon TB cases, compare the insights into transmission provided by genomic versus traditional epidemiological methods and determine the added value of molecular/genomic technologies in a setting with rich epidemiological data. (Manuscript 8)

4. To quantify the extent of local transmission of TB within BC over a ten-year period using genomic epidemiology, and demonstrate the utility of WGS through the reconstruction of likely transmission events within a large genomic cluster, while additionally characterizing all large clusters to reveal common trends in TB outbreaks. (Manuscript 9)
5. To address any remaining limitations of the previous manuscripts that could be investigated with available study data. (Manuscript 10)

1.3 Tuberculosis

1.3.1 Etiology

Tuberculosis is caused by the bacterium *Mycobacterium tuberculosis* (*Mtb*). When an individual with respiratory TB coughs, sneezes, or sings, droplets containing infectious bacilli become airborne.¹⁵ Following inhalation into the lungs, alveolar macrophages engulf the *Mtb* bacterium; however, *Mtb* has adapted to this hostile environment and are able to proliferate within the macrophage.¹⁶ Infected macrophages may carry bacilli to nearby lymph nodes, from whence they may migrate into the blood and disseminate throughout the body.¹⁷ Consequently, *Mtb* may infect multiple organs, or establish a localized infection in a particular organ or tissue.¹⁸ However, most cases of TB are confined to the lungs. Here, the immune system attempts to contain the bacteria by forming structures known as granulomas—an agglomerate of immune cells intended to “wall off” *Mtb* infection. Unsuccessful containment may support bacterial growth and lead to tissue necrosis, resulting in the formation of cavities in the lung tissue.¹⁹

There are three possible outcomes following *Mtb* exposure: bacilli clearance, latent TB infection (LTBI), or progression to active disease.²⁰ The frequency of successful clearance is unknown, and most exposures are assumed to result in LTBI—a clinically asymptomatic, non-infectious form of TB in which the bacilli has been successfully contained.²⁰ Characterized by the absence of clinical and radiographical signs of active disease, and an immune response detectable using a

tuberculin skin test (TST) or interferon gamma release assay (IGRA), LTBI is the most common manifestation of TB, representing 90–95% of infected individuals.^{21,22} The lifetime risk of progression to active TB disease amongst persons with LTBI is estimated at 5–10%, and the risk is highest within the first two years following infection.²² Active disease, which can develop within 1–2 months of exposure, is the potentially infectious form of TB, and symptoms include fever, chills, night sweats, weight loss and cough.¹⁸

1.3.2 Diagnosis and treatment

The tuberculin skin test, or TST, remains the most widely used diagnostic tool to determine whether a person has been infected with TB, though the test cannot distinguish between LTBI and active disease.²³ The test involves an intradermal injection of a small amount of tuberculin (an extract of the tubercle bacillus) into the inside forearm.²⁴ If a person has previously been exposed to a mycobacterial species resulting in an adaptive immune response, a swelling will appear at the site of injection. The size of the induration, if any, is measured 48–72 hours later by a healthcare professional. Various factors, such as prior BCG vaccine exposure, age, and immunocompromising conditions, will affect the size at which an induration is considered positive evidence for TB infection.^{25,26} In recent years, the IGRA blood test has also been used to diagnose LTBI. By evaluating the levels of interferon gamma produced in response to antigens specific to *Mtb* and not BCG or other non-tuberculous mycobacteria, IGRA may be preferred for diagnosing LTBI in certain populations.^{27,28} However, both IGRA and TST rely on an immune response and each has their own unique advantages and limitations.²³

For individuals with positive TST or IGRA results, a chest X-ray should be ordered to differentiate between latent TB infection and active pulmonary disease.⁶ Where there are radiographic signs of potential active TB disease, sputum specimens will be submitted for laboratory testing. Smear microscopy results indicating the presence or absence of acid-fast bacilli (AFB) are often the first test result obtained, with turnaround times of 1–2 days following receipt of the specimen by a specialized reference mycobacteriology laboratory. This test, if positive, provides an indication of bacterial load, scored from 1+ to 4+—the greater the number, the more bacilli are likely expelled and hence, the more infectious the individual.²⁹ Because AFB

microscopy is not specific for *Mtb*—other mycobacterial species will yield a positive result if present—a combination of molecular and culture-based methods are used to diagnose active *Mtb* infection. PCR assays to identify *Mtb* and *Mycobacterium avium complex*—a frequently observed non-tuberculous mycobacterium—may be performed on a smear-positive sample and give a result within days. All specimens, whether smear-positive or negative, are inoculated into culture, both in liquid and/or solid media; however, culture can take up to eight weeks to yield a positive result⁶ due to the lengthy doubling time of *Mtb* (~24 hours).³⁰ Upon culture confirmation, phenotypic drug susceptibility testing (DST) is recommended for the first culture-positive isolate from each new TB case to determine the most appropriate therapy.⁶ It should be noted that it is not always possible to obtain laboratory confirmation, particularly with respect to culture growth and some individuals will be clinically diagnosed with active TB.

Treatment of tuberculosis depends on whether the diagnosis is latent or active TB. Individuals with LTBI may be prescribed antibiotics that will substantially reduce the risk of progression to active TB in the future.³¹ However, the lengthy treatment time—typically six to nine months with one or more antibiotics—and the potential side effects of the drugs results in low uptake and completion rates.³² Estimates are that fewer than half of those initiating LTBI prophylaxis in North America complete the full regimen;³² however, new, shorter-course regimens requiring fewer and less frequent doses are now available and may improve completion rates. Treatment for active tuberculosis is not optional; however, the specific antibiotics and length of treatment may depend on an individual's age, comorbidities, anatomical site of infection, and possible *Mtb* antibiotic resistance.³³ Unless there is a high suspicion of multi-drug resistance (MDR)—defined as resistance to isoniazid and rifampin—initial empiric treatment using a combination of isoniazid, rifampin, pyrazinamide, and ethambutol is recommended for a minimum of six months.³³ Inadequate treatment of TB can quickly give rise to antibiotic resistance and poses a major global threat.^{34,35} Resistance is due to the acquisition of chromosomal mutations by *Mtb*; these mutations can be quickly detected with molecular tests such as PCR or line-probe assays, and large collaborations, such as Comprehensive Resistance Prediction for Tuberculosis: an International Consortium (CRyPTIC), are combining genomic data with phenotypic laboratory results to catalogue and discover new resistance-conferring mutations.³⁶

1.3.3 *Mycobacterium tuberculosis* lineages

The *Mycobacterium tuberculosis* complex is a genetically related group of *Mycobacterium* species and includes the human-adapted lineages *M. tuberculosis sensu stricto* and *M. africanum*, along with several lineages adapted to mammalian species.³⁷ Also included in the *Mtb* complex, *M. bovis* has the distinction of infecting both animals and humans. Close contact with infected animals and consumption of contaminated unpasteurized milk are the main routes of transmission to humans.³⁸

Genomic studies have enhanced our understanding of the global *Mtb* population structure, revealing a long history of host-pathogen co-evolution, following patterns of human migration resulting in a strong phylogeographic structure, meaning that genetically distinct lineages are associated with specific geographical regions.³⁹ *Mtb* has been classified into seven major phylogenetic lineages: lineage 1 (Indo-Oceanic), lineage 2 (East Asian), lineage 3 (East African-Indian), lineage 4 (Euro-American), lineage 5 (West African 1), lineage 6 (West African 2), and the recently identified lineage 7 found in Ethiopia or recent Ethiopian emigrants.^{40,41}

The genetic diversity across *Mtb* lineages manifests as phenotypic and epidemiological differences. For example, animal models have shown a more rapid progression to active disease and increased virulence for the East-Asian and Euro-American lineages.^{42,43} In addition, these two lineages are more globally widespread, are often responsible for large outbreaks, and are more likely to be associated with respiratory TB.^{37,44-46} It has been well established that East-Asian strains—specifically the Beijing sub-lineage—have a higher propensity for drug resistance.^{43,47} Overall, the characteristics of the pathogen are clearly linked to its ability to infect hosts, cause disease and spread from person-to-person.

1.3.4 Transmission

Various host, environmental, and pathogen factors influence the likelihood of TB transmission from an infected individual to another person.^{44,48,49} A host with active TB disease is the first element necessary for transmission. Whether or not an individual with active TB disease is infectious depends on multiple factors. Non-respiratory forms of the disease are assumed to be non-infectious, and the infectiousness of an individual with respiratory disease is influenced by their AFB smear status and the presence of cavitory disease,⁵⁰ although it should be noted that transmission from non-cavitory, smear-negative persons can occur,^{51,52} and precautionary measures, such as use of a mask or isolation, should be taken to minimize the possibility of transmission from all active TB cases until they are deemed non-infectious.⁶ Delays in diagnosis and treatment initiation also influence infectivity—they may increase the burden of disease and the length of time an individual is infectious, thereby increasing the number of social contacts that may result in TB transmission.^{53,54}

Beyond infectiousness of the source case, environmental factors, such as poorly ventilated indoor spaces,⁵⁵ duration of exposure, and physical proximity to the source case, may increase the probability of transmission.^{4,6,55} For this reason, congregate settings such as schools, long-term care facilities, prisons, and homeless shelters have been associated with TB outbreaks. Pathogen-level factors also play a role in transmission— as described in the previous section the East-Asian and Euro-American *Mtb* lineages demonstrate higher *in vitro* growth rates and virulence, and have been linked to higher numbers of secondary cases and more outbreaks in human populations, as compared to other lineages.^{37,44}

In addition to the factors listed above, a susceptible host is a further requirement for person-to-person transmission. Because the BCG vaccine does not effectively prevent infection—it instead prevents TB meningitis or disseminated disease in children exposed to TB⁵⁶—and because previous infection does not protect against reinfection,⁵⁷ all individuals are theoretically potentially susceptible to TB. However, evidence suggests that some individuals can clear TB infections despite exposure—for example, within a household, not everyone exposed to an infectious household member will become infected.⁵⁸ The proportion of close or household

contacts of an infectious source case that remain persistently TST-negative has been estimated at approximately half of those exposed.⁵⁹ While factors such as smoking,⁶⁰ substance use,^{61,62} malnutrition,⁶³ and immunocompromising conditions—particularly HIV⁶⁴—have been reported to increase susceptibility, other factors, including genetics, likely play a role in influencing an individual’s susceptibility.⁶⁵

The TB vaccine—Bacillus Calmette-Guérin (BCG)—was developed in 1921 and while it does provide some protection against severe forms of pediatric non-respiratory TB, it does not effectively prevent infection in adults nor does it prevent reactivation of LTBI.⁶⁶ Consequently, there is likely little to no impact of BCG vaccination on TB transmission rates, and a number of countries with low TB prevalence have discontinued universal BCG vaccination, including Canada.⁶

1.3.5 Monitoring, surveillance and investigation

Monitoring, surveillance, and investigation of active tuberculosis cases are essential elements of effective TB prevention and care programs. As a reportable disease, physicians in Canada are required to notify the appropriate local health authority of all new TB diagnoses. Each province independently maintains a provincial TB registry, which may be used to identify patterns and trends in TB epidemiology that inform planning around programmatic monitoring and disease surveillance initiatives.¹⁴ At the case level, local health authorities conduct contact investigations with the aim of identifying undiagnosed active cases and exposed individuals with LTBI in order to prevent the further spread of disease.⁶⁷ Establishing that a transmission between two individuals has occurred—particularly in an outbreak situation where there are multiple transmission events—is often possible only through interviewing newly diagnosed individuals to request the names of their close and casual contacts for TB screening.⁶⁸ Before the era of genomic epidemiology, identifying epidemiological linkages between cases through such interviews was considered sufficient for concluding that transmission had occurred; however, genomic studies have shown that these epidemiological assumptions may not always hold true.⁶⁹ Conversely, molecular and genomic methods have confirmed person-to-person transmission in instances where individuals were unwilling or unable to name contacts,^{70,71} and have also

revealed previously unsuspected routes of transmission.^{72–74} Given the potential for improving contact investigations and revealing patterns of local transmission, many low-incidence settings now, as standard practice, use molecular methods together with clinical and epidemiological information to identify transmission events.^{75–77}

1.3.6 Epidemiology of tuberculosis in Canada

In recent decades, the decline in TB incidence has plateaued in Canada, seeing little change from 2005 through 2014 (**Figure 1-1**). Over this period, an average of 1,634 active TB cases were reported each year, corresponding to an average ten-year incidence rate of 4.8 per 100,000 persons.⁷⁸ While this rate categorizes Canada as a low-incidence country—similar to the United States and many Western European countries⁷⁹—the rate is not uniform across the country (**Figure 1-2**). The Atlantic provinces have the fewest cases, together averaging 24 TB diagnoses annually (2005–2014), and with an incidence rate well below the Canadian average.⁷⁸ In contrast, 70% of Canada’s ~1,600 annual TB cases occur in Ontario, Quebec and British Columbia;⁷⁸ the metropolitan areas of Toronto, Montreal, and Vancouver account for most of these cases.^{8,80,81} The highest incidence rate for a province or territory is observed in Nunavut, where in 2012, the rate of 226 cases per 100,000 population was comparable to rates in sub-Saharan Africa.⁸²

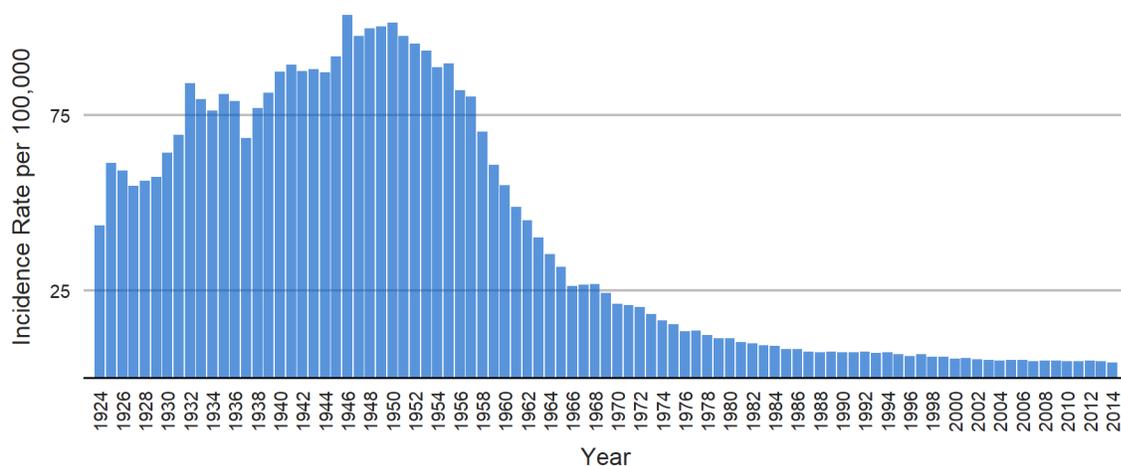


Figure 1-1. Tuberculosis incidence (per 100,000 population) in Canada, 1924–2014.

Data source: Public Health Agency of Canada;^{78,83} Created with: R (v3.4.1).

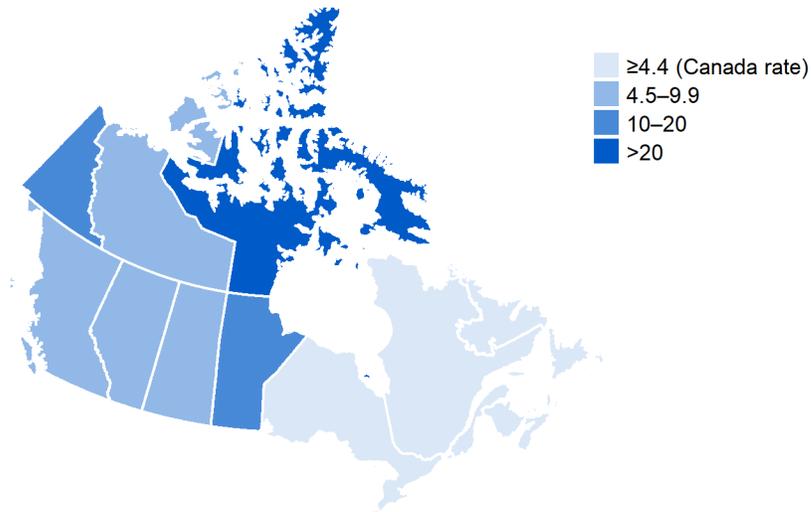


Figure 1-2. Tuberculosis incidence (per 100,000 population) by Canadian province/territory, 2014.
 Data source: Public Health Agency of Canada;⁷⁸ Created with: R (v3.4.1).

Persons born outside Canada make up the largest proportion of TB diagnoses overall, accounting for 70% of all active cases diagnosed in 2016.¹³ The proportion of TB cases in persons born outside Canada varies across the country, with the largest immigrant-receiving provinces (Ontario, Quebec and British Columbia) reporting the highest number of cases in migrants to Canada.⁶ Many of these cases are the result of LTBI reactivation from infections acquired outside Canada,⁶ and while migrants diagnosed with TB in Canada come from all nine World Health Organization regions, slightly more than half come from only four countries, all of which have high TB incidence rates: India, Philippines, China, and Vietnam.¹³

TB notifications differ by age and gender, with males accounting for a slightly higher proportion of TB cases (56%) compared to females (44%).¹³ With respect to age, persons 25 to 34 years of age proportionally represent the largest number of cases (18%).⁸⁴ However, TB rates in Canada are highest for those 75 years of age and older, at 10.4 cases per 100,000 population.¹³

Despite the global threat of multidrug-resistant TB,⁸⁵ rates of resistance remain low in Canada. Mono-resistance to any first-line antibiotic (isoniazid, rifampin, ethambutol, pyrazinamide) was detected in 7.5% of isolates and multi-drug resistance in 1.2% of isolates (2006–2010).⁶

1.4 Genotyping of *Mycobacterium tuberculosis*

Genotyping refers to a suite of molecular fingerprinting methods, in which specific DNA sequences are interrogated to identify genetically related isolates that may come from a common source. This ability to identify closely related TB isolates that might reflect recent endemic transmission is useful for population-level epidemiological studies, identifying person-to-person transmission, outbreak detection, identifying laboratory cross-contamination events, and differentiating between TB relapse and reinfection.^{86–93} Three tools are frequently used for *Mtb* strain typing: restriction fragment length polymorphism (RFLP), mycobacterial interspersed repetitive unit–variable-number tandem repeat (MIRU-VNTR), and spoligotyping.

1.4.1 RFLP

Fragmenting chromosomal DNA using restriction endonucleases and comparing the banding patterns that result from gel electrophoresis of the fragments is known as restriction fragment length polymorphism (RFLP) and is a commonly used technique in bacterial strain typing.⁹⁴ The RFLP protocol for *Mtb* leverages the insertional sequence *IS6110*, a 1,361-bp mobile element specific to the *Mtb* complex—this is detected via a labeled probe following gel electrophoresis.^{95,96} The number and location of *IS6110* elements varies (0–27 copies/genome),⁹⁷ resulting in a digital image of a banding pattern that can be used to compare isolates to each other. The estimated rate of evolution (3.2–8.7 years) of *IS6110* makes RFLP well-suited for epidemiological studies and outbreak detection,^{98–100} and it has the highest discriminatory power of the commonly used *Mtb* genotyping methods.¹⁰¹ However, the ability of RFLP to distinguish between strains is diminished in *Mtb* isolates with <6 copies of *IS6110*.^{77,102} Furthermore, RFLP is a labour-intensive method with a slow turnaround time, and the resulting gel images are difficult to compare between laboratories.⁹⁶

1.4.2 MIRU-VNTR

Two decades ago, 41 distinct mini-satellite-like structures were identified across the *Mtb* genome.^{103,104} Composed of 40–100-bp variable number tandem repeats (VNTRs), these structures, located mainly in intergenic regions, are known as mycobacterial interspersed

repetitive units (MIRUs).¹⁰³ MIRU-VNTR is a PCR-based method used to determine the number of these repetitive units at specific loci.¹⁰³ The technique involves PCR amplification using fluorescently labeled primers complementary to flanking DNA regions, and sizing of the resulting amplicons using an automated capillary electrophoresis DNA analyzer.¹⁰⁵ Software such as GeneMapper (Applied Biosystems)^{77,106} or BioNumerics (Applied Maths)¹⁰⁷ are used to determine the number of repeats, and the final result is a concatenated string of digits representing the number of repeated sequences at each loci analyzed (e.g. 223454124341). A locus with >9 repeats is represented by a letter code, A=10, B=11, C=12, etc. The discriminatory power of MIRU-VNTR depends upon the number of loci typed. Initially, 12 loci were used, which was then expanded to 15, and currently, a set of 24 loci have been agreed upon as the optimal scheme that balances maximal strain discrimination against the time and costs associated with genotyping.⁷⁷ 24-locus MIRU-VNTR is currently considered the global standard for TB genotyping, and is performed by most well-resourced reference mycobacteriology laboratories.¹⁰⁸ Due to the low amount of input DNA needed, the rapid turn-around time, and the portable digital pattern generated, MIRU-VNTR has surpassed RFLP as the preferred method of genotyping.^{102,108}

1.4.3 Spoligotyping

Another commonly used PCR-based *Mtb* genotyping method is spacer oligonucleotide typing, otherwise known as spoligotyping. This technique makes use of polymorphisms in the direct repeat (DR) locus, in which a multiple, highly-conserved 36-bp DRs are interspersed with 35–41bp spacers.¹⁰⁹ Following PCR amplification of the DR region, amplicons are hybridized to a set of 43 spacer oligonucleotides—the presence of each is detected by a membrane-based reverse line-probe assay or, more recently, using a microbead system.¹¹⁰ A binary code (1–0) is used to indicate the presence or absence of each of the 43 spacers, which is then converted to a more compact 15-digit octal code (e.g. 777774777303761).¹¹¹ Spoligotyping's advantages are similar to those of MIRU-VNTR—a low amount of input DNA is needed, the method has a rapid turnaround time, and it results in a digital code easily comparable between laboratories; however, spoligotyping has considerably lower resolution compared to RFLP and MIRU-VNTR and therefore is most useful when paired with one of these higher-resolution techniques.^{111–113}

1.5 Molecular Epidemiology and Public Health

The use of molecular biology-based genotyping assays in the context of traditional epidemiological investigations is known as molecular epidemiology, and it greatly improved our understanding of the *Mtb* population structure both within specific regions and globally.^{41,114–116} Genotyping facilitates the identification of clusters of TB isolates potentially reflecting recent transmission, and has improved public health’s ability to detect outbreaks and identify person-to-person transmission events, when used in combination with traditional contact investigation methods.^{74,117,118} Molecular technologies have also proven useful in distinguishing between relapse and exogenous reinfection—key to providing insight into the biology of reoccurrences and understanding treatment failure.^{119,120} At a population-level, molecular epidemiology has allowed for a more accurate quantification of disease epidemiology, particularly around the estimation of cases attributable to LTBI reactivation versus local transmission.^{121–124}

1.6 A Brief Primer on Genomic Epidemiology: Lessons Learned from *Mycobacterium tuberculosis*

1.6.1 The *Mycobacterium tuberculosis* genome

In 1998, the Wellcome Trust Sanger Centre sequenced the first *Mycobacterium tuberculosis* (*Mtb*) genome—the laboratory strain H37Rv.¹²⁵ That was followed in 2002 by the genome of CDC1551, a clinical isolate sequenced at the Institute for Genomic Research.¹²⁶ Since then, dramatic improvements in sequencing technology and capacity have led to a wealth of *Mtb* genomes. The National Center for Biotechnology Information (NCBI) now stores over 3,600 assembled *Mtb* genomes¹²⁷—47 are considered complete, with no gaps or ambiguous bases, and many thousands are partially assembled. The number of unassembled *Mtb* genomes is even larger—NCBI’s Sequence Read Archive stores raw read data from ~22,000 isolates,¹²⁸ with thousands more data sets deposited each year.

It is no surprise that so many *Mtb* genomes have been sequenced. In its most recent Global Tuberculosis Report,¹²⁹ the World Health Organization reported that, in 2014, 9,600,000 people had active tuberculosis (TB) disease, with 5% of those cases demonstrating multidrug resistance—resistance to two first-line antibiotics, isoniazid (INH) and rifampin (RIF). As much as one third of the world’s population is thought to harbor latent TB infection,¹³⁰ with between 5% and 10% of these individuals progressing to active, symptomatic disease at some point in their lives.²²

Beyond the scale of the problem, the nature of the *Mtb* genome itself also promotes genomic inquiry. *Mtb* is a highly clonal species.¹³¹ It comprises seven geographically structured lineages, with a maximum difference of only 1,800 single nucleotide variants (SNVs) between lineages.⁴⁴ Evolution has largely proceeded through genomic deletions, with minor historical contributions from repetitive and mobile genetic elements (MGEs).⁴⁴ The obligate intracellular lifestyle of *Mtb* means that horizontal gene transfer is rare,¹³² as is recombination.¹³³ *Mtb* harbors no plasmids, and antibiotic resistance within *Mtb* arises solely through chromosomal mutations.¹³⁴ Together,

these factors make the assembly and analysis of *Mtb* genomes a comparatively straightforward proposition.

1.6.2 Leveraging genomics for epidemiology

Managing TB, particularly in low- and medium incidence settings, such as North America and Europe, relies on several strategies, from rapid diagnostics and drug sensitivity testing to removing barriers around accessing care and adhering to treatment.¹³⁵ In addition to appropriate therapy, one of the most fundamental strategies for TB management is epidemiological investigation. The guidelines for investigation were first set out by the American Thoracic Society in 1976,¹³⁶ and were updated in 2005.¹³⁷ At its simplest, investigation involves interviewing a newly diagnosed individual to establish a list of their named contacts and places of social aggregation, followed by contact investigation—prioritizing contacts for follow-up and performing a tuberculin skin test (TST) or another screening instrument. At its most complex, an investigation might involve screening thousands of individuals in large institutional settings or across multiple geographic areas and instituting a multiyear outbreak-management plan.

TB outbreak investigations have been greatly enhanced by molecular epidemiology methods, reviewed by Kato-Maeda *et al.*¹⁰¹ and illustrated in **Figure 1-3**. By examining specific regions of the *Mtb* genome independently or in combination, including IS6110 insertion elements, the CRISPR direct repeat locus, or 12–24 mycobacterial interspersed repetitive unit (MIRU) variable number tandem repeat (VNTR) loci—TB laboratories can determine whether two or more *Mtb* isolates have the same genotype, suggesting recent transmission. Prospective genotyping of all culture-positive *Mtb* isolates¹³⁸ is now routine practice in many well-resourced mycobacteriology laboratories,¹³⁹ facilitating the real-time identification and investigation of TB clusters.

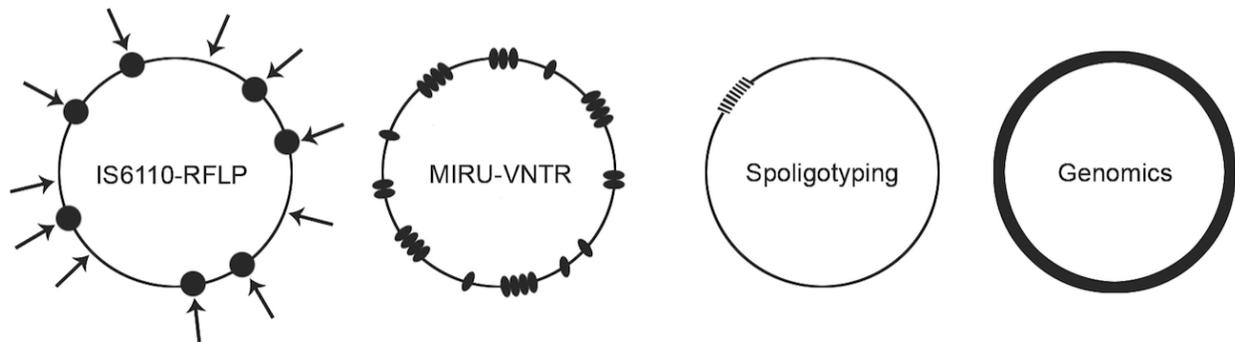


Figure 1-3. Molecular epidemiology methods used in tuberculosis surveillance. This simple schematic, not to scale, compares four common methods used in TB molecular epidemiology. In *IS6110*-RFLP, a restriction enzyme cuts the genome (shown as arrows) and the resulting fragments are separated using gel electrophoresis. A probe specific to the *IS6110* insertion element (shown as dots), which varies in count and position between isolates, is applied, resulting in a distinct banding pattern. In MIRU-VNTR, PCR is used to amplify between 12 and 24 VNTR loci (ovals) in the genome, and the products are separated by either capillary or gel electrophoresis alongside a size standard to determine the length of the amplicon, allowing for calculation of the number of repeats, which is then converted to a digital code to facilitate comparison against a database. In spoligotyping, a hybridization assay is used to detect the presence or absence of 43 “spacer oligonucleotides” in the direct repeat region (hatched lines), a pattern that is translated into a binary representation and then an octal code. Unlike the other methods, genomics interrogates the entire genome, with SNVs revealing the relationship between isolates.

Despite the many insights into TB transmission afforded by molecular epidemiology, there are several attendant limitations. Its resolution depends on which typing method is used, giving varying answers as to the size and membership of clusters—large clusters described by 12-locus MIRU-VNTR often collapse into multiple, smaller clusters when 24-locus MIRU-VNTR was implemented. Genotyping cannot definitively rule in transmission, though it can rule it out,¹³⁹ and it often overpredicts clustering in certain lineages and sublineages.¹⁴⁰

Given that genotyping only interrogates a fraction of the available genomic information in an isolate, an obvious question to ask is “what would happen if we looked at the whole genome?” Because outbreak investigations often involve tens, if not hundreds, of individual bacterial isolates, this notion of genomic epidemiology was simply not possible using the Sanger sequencing-based approaches that generated the first *Mtb* genomes. It was only with the release

of second-generation sequencing platforms,¹⁴¹ with their high throughput and low per-genome cost, that microbial genomics entered the era of large-scale sequencing.¹⁴²

The concept behind genomic epidemiology is simple. As bacteria replicate, mutations arise. Over the short time scale of an outbreak and absent selective pressures, these accrue in a largely neutral fashion according to a molecular clock. If a mutation is present in the genome of a pathogen infecting person A, that mutation will also likely appear in the genomes of the pathogens isolated from everyone person A has infected. By reading the complete genomes of the pathogens isolated from each case in an outbreak, the patterns of shared mutations suggest transmission events (**Figure 1-4**).¹⁴³ It is important to note that, although methods are presently being developed to infer transmission directly from phylogenetic trees derived from whole-genome data,^{144–146} it is only by assessing the transmission events suggested by genomics in the light of available epidemiological data that investigators can draw reliable conclusions about who infected whom.¹⁴⁴ Drawing further conclusions around transmission, such as where and how it might have occurred, will always require comprehensive epidemiological data—in a nosocomial outbreak, for example, genomic data from patient and environmental isolates might suggest the “who” of transmission, but it is only through examining patients’ movements and their exposure to common environments or equipment that the “why” and “how” become clear and can inform infection-control activities.

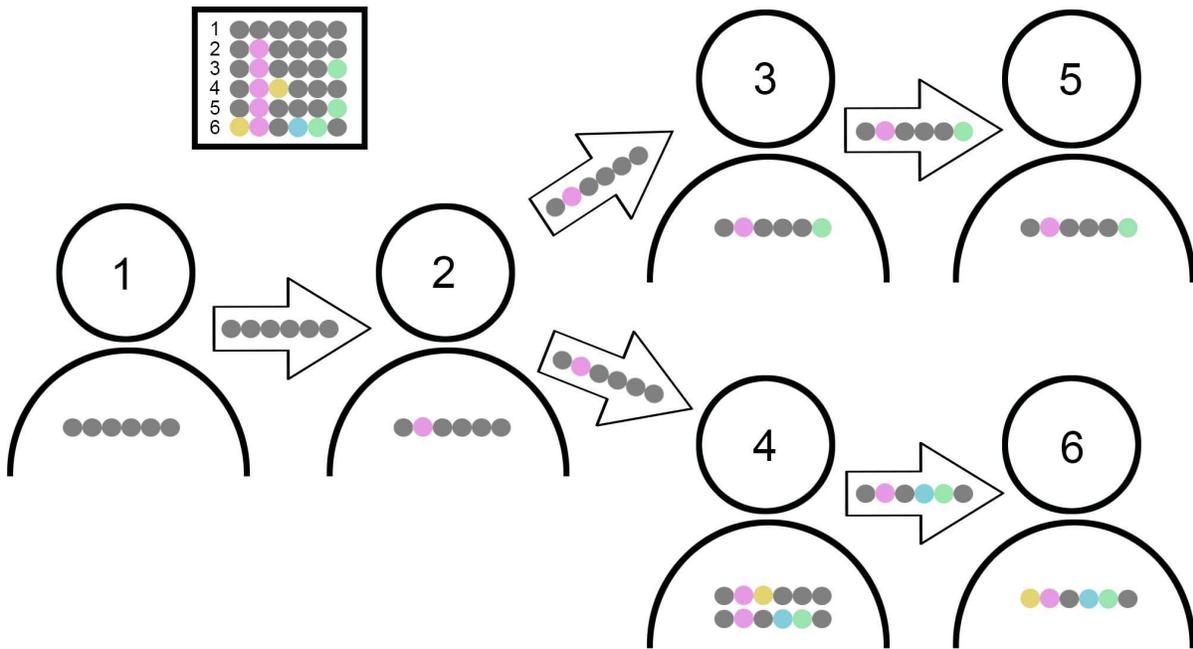


Figure 1-4. The basic principle of genomic epidemiology. In this small transmission tree, six infected individuals, or hosts, are shown. Each host is infected with a simplified “pathogen,” shown here with a simple 6 base pair genome, where gray circles indicate a wild-type base and colored circles represent different mutations (A, C, G, or T). Host 1 transmits their pathogen to host 2, in whom a single mutation arises (pink); this mutation is seen in all cases downstream of host 2. Host 2 infects host 3 and 4. In host 3, a second mutation (green) arises and is passed to host 5, in whom no further mutations accrue. Host 4 is an example of within-host diversity. Owing to a prolonged and/or disseminated infection or a burst of mutation, host 4 harbors a pool of variation, with the infecting strain from host 2 as a common ancestor. Host 4 transmits one of their lineages to host 6. The inset panel shows the results of sequencing the pathogens sampled from each host; these data can be combined with clinical and epidemiological data to reconstruct likely transmission pathways.

Some of the earliest works in genomic epidemiology focused on TB. In 2009, Schürch *et al.*¹⁴⁷ sequenced the first and last isolates from a known five-person transmission chain, identifying six variants, which they surveyed in isolates from persons 2, 3, and 4 using amplicon sequencing. They found that the bulk of the variants arose in a single host—person 4—who was noncompliant with treatment. In 2011, the Gardy laboratory sequenced 36 *Mtb* genomes from a large outbreak in British Columbia, Canada.⁷⁴ By combining the genomic data with epidemiological information collected through a social network questionnaire,¹⁴⁸ plausible transmission events were inferred for most cases in the outbreak and identified super-spreaders who contributed disproportionately to infection of secondary contacts. In 2012, Walker and colleagues scaled up the genomic epidemiology approach by an order of magnitude, sequencing 390 isolates from 254 cases.¹⁴⁹ Their work provided several important benchmarks, establishing a mutation rate of 0.5 SNVs per year and suggesting thresholds of five SNVs as a maximum distance between epidemiologically linked cases and 12 SNVs as a threshold above which transmission can be ruled out. These thresholds are unique to TB and its slow mutation rate; for other pathogens with different molecular clocks and shorter outbreaks, other thresholds will apply. Duchêne *et al.*¹⁵⁰ recently calculated evolutionary rates for 35 human pathogens, ranging from the very low (*M. tuberculosis* at 5.39×10^{-8} substitutions per site per year, or ~ 0.3 SNVs per year) to the very high (vancomycin-resistant *Enterococcus faecium* at 9.35×10^{-6} substitutions per site per year, or ~ 26 SNVs per year).

Since those first studies, the use of genomics to understand TB transmission has dramatically increased, with 27 papers and over 5000 *Mtb* genomes to date (**Table 1-1**). Genomic epidemiology is also being applied in the larger clinical microbiology space, with studies of everything from food- and water-borne pathogens to hospital-acquired infections, as highlighted in several recent reviews.^{143,151–154} Pathogen genomics has become an integral part of many national public health agencies' routine practice, with groups, such as Public Health England (PHE), the U.S. Centers for Disease Control and Prevention (CDC), and the Public Health Agency of Canada (PHAC), all engaged in genomic surveillance.

Though the technique is now being applied across the microbial landscape, many of the lessons learned around genomic epidemiology arose from work in TB. This review will provide a brief primer on genomic epidemiology, highlighting lessons learned from *Mtb* that have broader implications for work in other organisms.

Table 1-1. Genomic epidemiology TB studies examining transmission between individuals.

Publication	<i>Mtb</i> isolates sequenced
Schürch <i>et al.</i> ¹⁴⁷	2
Schürch <i>et al.</i> ¹⁵⁵	3
Gardy <i>et al.</i> ⁷⁴	36
Bryant <i>et al.</i> ⁶⁹	199
Bryant <i>et al.</i> ¹¹⁹	94
Clark <i>et al.</i> ¹⁵⁶	51
Kato-Maeda <i>et al.</i> ¹⁵⁷	9
Roetzer <i>et al.</i> ¹⁵⁸	86
Török <i>et al.</i> ¹⁵⁹	2
Walker <i>et al.</i> ¹⁴⁹	390
Jamieson <i>et al.</i> ¹⁴⁰	36
Kohl <i>et al.</i> ¹⁶⁰	26
Luo <i>et al.</i> ¹⁶¹	32
Mehaffy <i>et al.</i> ¹¹⁷	61
Pérez-Lago <i>et al.</i> ¹⁶²	36
Walker <i>et al.</i> ⁷³	247
Coscolla <i>et al.</i> ¹⁶³	46
Glynn <i>et al.</i> ¹⁶⁴	170
Guerra-Assunção <i>et al.</i> ¹⁶⁵	1687
Guerra-Assunção <i>et al.</i> ¹²⁰	1933
Lee <i>et al.</i> ¹⁶⁶	120
Lee <i>et al.</i> ¹⁶⁷	163
Regmi <i>et al.</i> ¹⁶⁸	9
Stucki <i>et al.</i> ¹⁶⁹	3
Witney <i>et al.</i> ¹⁷⁰	16
Arnold <i>et al.</i> ¹⁷¹	2
Outhred <i>et al.</i> ⁷²	22

1.6.3 Step 1: look before you leap

With the ever-decreasing cost of bacterial genome sequencing—a recent review noted that, in some extreme cases, the cost of generating a single draft assembly is now less than one dollar¹⁷²—it can be tempting to see sequencing as the hammer for every nail. However, before deciding to launch a genomic epidemiology study, several questions must be answered.

First, is sequencing even possible? To infer transmission events, one must have a well-sampled outbreak; in other words, DNA should be available for the majority of the outbreak cases. For a reportable disease like TB, with nearly exclusive human-to human transmission, and in a well-resourced, low incidence setting where culture is routinely used in diagnosis, this is often achievable. For example, in a recent British Columbia TB outbreak,¹⁷³ *Mtb* DNA was obtained from 48 of 52 outbreak cases.¹⁷⁴ For non-notifiable diseases and/or those involving food- or water-borne transmission or nonhuman hosts, the outbreak may be insufficiently represented by the available samples.

A further wrinkle is introduced by the nature of the diagnostic test used. For certain pathogens—largely viruses but also bacteria, such as *Treponema pallidum* or *Streptococcus pyogenes*—diagnosis is often based on serology and the pathogen is never cultured. In the recent work on a British Columbia measles outbreak, 45% of the cases were diagnosed on clinical or serological grounds and did not have available nucleic acid, precluding the inference of person-to-person transmission events beyond those already suggested by the field epidemiology.¹⁷⁵ Even the use of molecular diagnostics is no guarantee that genomic DNA will be available. As clinical microbiology enters a new era in which the focus is increasingly on rapid and direct molecular identification of pathogens without culture through platforms such as MALDI-TOF,¹⁷⁶ we may lose the opportunity to extract DNA in sufficient quantities for sequencing.

Nevertheless, the microbial genomics community is rising to meet these challenges. Several groups have developed library preparation methods capable of sequencing from ultra-low-input samples—on the order of picograms of nucleic acid,¹⁷⁷ while others are working toward extracting pathogen DNA directly from clinical samples, including urine,¹⁷⁸ vaginal swabs,¹⁷⁹

and even sputum.¹⁸⁰ Even more promising is the notion of clinical metagenomics—directly sequencing all of the nucleic acid present in a sample and using the resulting data to both diagnose an infection and assemble a genome that can then be scanned for resistance-associated mutations and/or epidemiological markers. This was first demonstrated by Loman *et al.*¹⁸¹ in their investigation of an *Escherichia coli* O104:H4 outbreak, with subsequent proof of principle in *Mtb* sequencing from sputum.¹⁸² Clinical metagenomics was recently reviewed in the context of both Salmonella diagnostics,¹⁸³ and microbial infections in general.¹⁸⁴

Assuming there is DNA to be sequenced, can you afford it? The National Human Genome Research Institute estimates the costs of sequencing, including production costs, at \$0.014 per megabase,¹⁸⁵ data that give rise to estimates like the previously described sub-\$1 bacterial genome.¹⁷² Unfortunately, these figures can lead to sometimes unrealistic expectations of the costs of a bacterial genomics project. Such economies are possible in large sequencing centers, but for the typical laboratory outsourcing sequencing to a third-party provider or using commercial reagent kits, is likely the cost will be closer to ~\$100–\$250 per bacterial genome, not including upstream laboratory costs for culture and extraction and downstream personnel costs for bioinformatics and interpretation.

This notion of interpretability is also important to consider from multiple perspectives. First, is the genome itself readily interpretable in an epidemiological context? As noted earlier, *Mtb* has a single nonrecombining chromosome, no plasmids, and a well-described mutation rate.¹⁸⁶ In highly recombinogenic pathogens, such as *Neisseria meningitidis*, *Streptococcus pneumoniae*, and *Campylobacter jejuni*, phylogenetic trees—the foundations of many genomic epidemiology analyses—may be distorted. Recent work has shown that, although the topologies of such trees are robust to recombination, branch lengths and the parameters derived from them—data integral to making epidemiological inferences—are not.¹⁸⁷ Important work on the genomic epidemiology of hospital-associated outbreaks of carbapenemase-producing organisms has demonstrated that it is not just strains that move between persons and the environments, but also plasmids,¹⁸⁸ and that this plasmid movement depends on a range of factors.¹⁸⁹ Reconstructing plasmids from short-read data requires additional bioinformatics steps,^{190,191} and, even then, the acquisition of

exogenous genes through plasmids, integrons, and transposons often complicates the identification of transmission chains.¹⁹² At the chromosomal level, mutation rates can vary substantially between pathogens,¹⁹³ meaning that a simple threshold for the number of SNVs associated with chains of transmission cannot always be determined.¹⁹⁴

The second issue around interpretability is whether or not a research group has the capacity for the downstream bioinformatics and epidemiological interpretation. Our experiences have shown that, in order to extract meaningful, actionable information from a genomic epidemiology study, it is not enough just to be able to run the software—instead, the person interpreting the resulting data needs to understand the disease and its epidemiology and appreciate the unique aspects of a pathogen’s genome. In *Mtb*, for example, repetitive and low-complexity regions make up about 10% of the total genome and are typically discarded from a bioinformatics analysis owing to difficulties in accurately calling SNVs in these regions.¹⁹⁵ If a researcher leaves these regions in his or her analyses, they can easily be misled by assembly errors masquerading as mutations, leading to incorrect assumptions around mutation rate and transmission events. A similar caution applies to the recombinogenic pathogens—if a researcher does not understand their pathogen’s genomic quirks and account for them in their bioinformatics analysis, the wrong conclusions may be drawn. Unfortunately, finding a team member with the right combination of epidemiological, computational, and evolutionary biology expertise is extremely difficult, and many groups struggle to even find someone with just the computational background. For this reason, a number of publicly available, pathogen-specific assembly and variant-calling platforms are being released. Tools like TGS-TB for *Mtb*¹⁹⁶ and WGSAnet for *Staphylococcus aureus*¹⁹⁷ facilitate bioinformatics analyses and incorporate the organism-specific knowledge necessary to provide high-quality output that users can have confidence in.

Data access is yet another issue to consider. Whether sequencing outbreaks or using genomics as a tool to understand epidemic dynamics,^{198–200} one must ask whether clinical and epidemiological case-level data are available (e.g., symptom onset, infectious period, named contacts, locations, hospitalization dates), and whether it can be linked to specimen-level results, such as phenotypic drug susceptibility testing (DST) data.

Finally, it is helpful to ask a somewhat difficult question—why use genomic epidemiology? Genomic epidemiology is effective when there is a clear outcome with public health relevance, such as implementing infection control procedures,^{201,202} identifying the source of infection,²⁰³ changing outbreak management guidelines,^{68,204} and creating surveillance resources to support prospective sequencing efforts.²⁰⁵ Where other interventions are better suited to affecting a disease's epidemiology—particularly in the developing world, where diagnosis, access to care, and adherence to treatment are fundamental issues—we would argue that the resources required for a genomic epidemiology investigation might be better used to address these more foundational issues.

1.6.4 Step 2: from sample to sequence

Having decided to launch a genomic epidemiology investigation, the next step is to sequence each sample of interest. As noted earlier, efforts are underway to sequence directly from clinical samples; however, most genomic epidemiology investigations currently begin with an isolate in pure culture. While certain sequencing services accept cultures under some circumstances, sequencing typically proceeds from extracted DNA. The choice of extraction method depends on the nature of the underlying sample,²⁰⁶ employing either in-house protocols, such as ethanol or phenol–chloroform extractions, or commercial kits—either low-throughput kits or high-throughput automated handling systems, such as the MagMAX (ThermoFisher, Waltham, MA) or QIAasympyphony (Qiagen, Germany). Extraction may be followed by enrichment, in which a bait is used to capture a specific region of interest²⁰⁷ or remove nonbacterial data from a metagenomic sample,²⁰⁸ or in which a whole-genome amplification technique is used to increase the amount of nucleic acid available for sequencing.²⁰⁹ Extracted DNA is then quantified using a platform such as the Qubit, Nanodrop, or Quant-iT (ThermoFisher, Waltham, MA), though results across platforms tend to vary, and the reported concentrations depend heavily on sample quality.^{210,211} DNA may also be checked for integrity using the Bioanalyzer or TapeStation (Agilent, Santa Clara, CA); degraded samples are likely to yield poor results downstream. Once each sample has been quantified, a decision can be made as to which to move forward to library preparation. Typically, sequencing centers will request at least 500 ng of total DNA, even though certain library preparation kits, such as Illumina's Nextera XT, are capable of sequencing from

as little as 1 ng of starting material. Both Head *et al.*²¹² and van Dijk *et al.*²¹³ recently reviewed library preparation methods, discussing the range of approaches for fragmentation, size selection, and adaptor ligation. Unlike library preparation methods for more specialized applications (e.g., *de novo* assembly and finishing of a new genome, RNAseq, ChIP-seq, or methylation studies), issues such as low-concentration starting material or GC-biased genomes do not confound most resequencing studies, in which the resulting reads will be aligned against an existing reference genome. Indeed, for the British Columbia *Mtb* resequencing work, in which over 1,500 complete genomes have been sequenced to date, excellent results have been achieved using both commercial kits and in-house protocols developed by a local sequencing center, with high-quality data from as little as 2–3 ng starting material.

Experience suggests that the most critical part of step 2 is selecting a sequencing platform and a multiplexing scheme. Too little sequencing capacity and/or over multiplexing will result in low-coverage, unusable data, while over sequencing is a poor use of resources, and the resulting files can be too large to manage and transfer efficiently.

Table 1-2 briefly summarizes currently available sequencing platforms, their throughputs, and the advantages and disadvantages of each system in the context of a genomic epidemiology study. Thanks to their low error rates and large market penetration, the Illumina platforms dominate the genomic epidemiology market space, with national public health agencies relying on networks of MiSeqs^{214,215} or centralized HiSeqs²¹⁶ for their genomic diagnostic and surveillance efforts. Short-read data are frequently analyzed in tandem with long-read data from the PacBio platforms to elucidate plasmid transmission in hospital-associated infections,¹⁸⁸ and proof of principle for plasmid sequencing on the Oxford Nanopore was recently demonstrated by PHE in their investigation of an outbreak of ST38 *E. coli* with a chromosomally integrated blaOXA-48 element.²¹⁷ Use of the Ion platforms and the Oxford Nanopore MinION is more common in the viral genomic epidemiology space, with the notable exception of the real-time nanopore sequencing and management of a hospital *Salmonella* outbreak.²¹⁸ As nanopore technology matures—the first MinION data was only released in 2014²¹⁹—this is likely to change.

Table 1-2. Sequencing platforms currently deployed for genomic epidemiology studies .

Platform	Throughput	Platform Throughput Comments ²²⁰
Illumina MiniSeq ²²¹	7.5 Gb, 25 M reads Max 2×150 bp 4–24 h run time	High accuracy reads (0.1% error rate), ideal for identifying variants for a genomic epidemiology study. Range of platforms offered, capable of handling small projects (MiniSeq, MiSeq) to large-scale efforts (HiSeq). Can be coupled to Neoprep automated library prep platform.
Illumina MiSeq ²²¹	15 Gb, 25 M reads Max 2×300 bp 4–55 h run time	
Illumina NextSeq ²²¹	120 Gb, 400 M reads Max 2×150 bp 12–30 h run time	
Illumina HiSeq ²²¹	1500 Gb, 5 B reads Max 2×150 bp 1–6 days run time	
Ion PGM ²²²	2 Gb, 400 k–5.5 M reads 200 or 400 bp 2–7 h run time	Simple machine—less subject to breakdowns. Low (1%) error rate.
Ion S5/S5 XL ²²³	1–15 Gb, 3–80 M reads 200 or 400 bp 3–18 h run time	Can be coupled to Ion Chef automated library prep platform.
Ion Proton ²²⁴	10 Gb, 80 M reads Up to 200 bp 2–4 h run time	
Oxford Nanopore MinION ²²⁵	Up to 42 Gb, 4.4 M reads Up to 300 kbp reads reported 1 min–48 h run time	Small, fast, and portable. Higher error rates (4–12%). ²²⁶ Not yet widely used in bacterial genomic epidemiology.
PacBio RSII ²²⁷	1–16 cells, 150 k reads/cell Up to 60 kbp reads reported 30 min–6 h run time	Short run time. Good for sequencing plasmids. Often used in tandem with Illumina short-read data.
PacBio Sequel ²²⁷	1–16 cells, 1 M reads/cell Up to 60 kbp reads reported 30 min–6 h run time	

To make the most efficient use of a sequencer's throughput, it is almost always necessary to multiplex samples. By ligating unique barcodes to DNA fragments during the library preparation stage, as many as 384 samples can be pooled, run in a single sequencing lane, and separated bioinformatically after sequencing. The degree to which one can multiplex a batch of samples depends on both the sequencing platform and the desired coverage—the number of times each base pair in a genome is sequenced. The microbial genomics community has largely gone with a consensus of 50× coverage,²²⁸ and, by using a coverage calculator, one can easily calculate how many samples can be multiplexed on a single run.²²⁹ In our own work on *Mtb* in BC, typically no more than 15 genomes are multiplexed on a single 2×150 bp MiSeq run using the v2 reagent kit (Illumina, San Diego, CA), giving ~60× coverage. When outsourcing sequencing to a local sequencing center, their maximum indexing capacity is 92 samples in a single HiSeq lane, which gives us an average coverage of ~150×.

1.6.5 Step 3: bases to bytes

Whether sequencing is done in-house or outsourced, the resulting data will almost always be returned as FASTQ files, which store both the sequence of each read and a quality score for each base in the read.²³⁰ In order to go from FASTQ files to a set of isolate-specific mutations that can be used to infer transmission events, there are a series of bioinformatics steps—alignment, variant calling (where “variant” means both SNVs and insertions or deletions, collectively referred to as indels), and variant filtering—that must be executed, either via a series of command-line tools or a pre-packaged, organism-specific pipeline. A full discussion of the command-line tools is outside the scope of this review; however, an excellent tutorial on genome assembly and analysis was published by Edwards and Holt in 2013,²³¹ and updated in a 2016 blog post.²³² The original publication describes the assembly, annotation, comparative analysis, and typing of an *E. coli* O104:H4 genome and, in a supplementary file, provides detailed instructions so that the reader can install, run, and understand the output of the many tools highlighted in the tutorial.

Alignment

Genomic epidemiology studies are resequencing studies, in which multiple isolates of a well characterized species, like *M. tuberculosis*, *E. coli*, or *Staphylococcus aureus*, are sequenced and compared to each other to identify the point mutations or other changes that suggest evolutionary and epidemiological relationships among the samples. In these scenarios, short reads are aligned against a high quality, completely finished reference genome to identify regions of difference—an approach known as reference mapping or reference-guided assembly. This stands in contrast to *de novo* assembly, in which overlapping short reads are slowly built up into long stretches of contiguous sequence called contigs without the use of a reference genome.²³³ *De novo* assembly is typically used for comparative genomics studies, in which many spatially and temporally diverse isolates are sequenced in order to understand a species' population structure or to identify virulence factors and antimicrobial resistance genes, or in outbreak studies of an emerging pathogen with few reference genomes, as in the case of the recent *Elizabethkingia anophelis* outbreak.²³⁴

In pursuing a reference-mapping strategy, experience with *Mtb* has suggested a number of best practices. First, recent work from Lee and Behr²³⁵ demonstrated that, when the goal of a resequencing study is to identify transmission-informative SNVs, the choice of reference genome in a clonal species like *Mtb* does not matter. While mapping a set of reads to references from different lineages does result in varying numbers of total SNVs identified, when these total SNV lists are filtered to leave only those SNVs that vary between outbreak isolates, the final count is the same regardless of the reference chosen. In less clonal species, a fast, hash-based distance estimation tool like Mash²³⁶ should be used to compare a set of reads to the NCBI RefSeq database and to select an appropriate reference.

Second, while initial outbreak investigations used the CDC1551 reference genome, selected for its similarity to an outbreak strain in BC, we have since changed our protocol to map all data against the H37Rv reference for the simple reason that it facilitates interlaboratory comparisons. Nearly every study in **Table 1-1** has used H37Rv as its reference, and, in their prospective COMPASS-TB project, PHE selected H37Rv as the reference against which their *Mtb* reads are

aligned.²¹⁴ Furthermore, many of the *Mtb* databases warehousing resistance associated SNVs, reviewed by Stucki and Gagneux,²³⁷ list SNVs according to their position in the H37Rv reference.

Finally, it is helpful to perform a *de novo* assembly of any unmapped reads remaining after reference mapping—this typically results in at least 5 to 10 contigs of length >1 kb. These regions, which are small stretches of coding sequence present in the outbreak isolates but not the H37Rv reference, occasionally contain informative SNVs. A recent review by Mielczarek and Szyda²³⁸ summarizes the tools used in reference mapping and their underlying algorithmic principles; of the many methods they describe, BWA²³⁹ is used by the majority of the microbial genomics community. In our own *Mtb* work in BC, we run the BWAmem algorithm, the most recent implementation of BWA. Before alignment, FastQC²⁴⁰ is used to verify the quality of short-read data and Trimmomatic²⁴¹ to remove sequencing adaptors and bases with a low-quality score. While trimming is not strictly necessary when using BWAmem, it does improve the performance of other aligners. The output of the alignment stage is typically a Binary Alignment Map (BAM) file, which can be further refined using an assembly improvement tool, such as Pilon,²⁴² or the local realignment workflow in GATK.²⁴³

Variant calling and filtering

After reference mapping, the next step is to identify variants present in one's samples but not in the reference genome. There are many approaches to variant calling; a full description of these is provided by Mielczarek and Szyda.²³⁸ Pabinger *et al.*²⁴⁴ offer a similarly detailed review of the many tools available; of these, both SAMtools mpileup²⁴⁵ and GATK²⁴³ are widely used in genomic epidemiology. These tools take a BAM file as input and output a list of SNVs and indels as a VCF (variant call format) file,²⁴⁶ in which the genomic coordinates of each variant are provided, along with columns describing the genotype at that position, the read depth, the genotype quality, and many other metrics. Arguably more important than the choice of variant caller is the approach one uses to filter the resulting VCF file, which will invariably contain a number of false-positive SNVs. While a small fraction of errors are introduced during sample preparation or sequencing—these can often be corrected with tools, such as Blue²⁴⁷ or

QuorUM²⁴⁸—most SNV-calling errors result from improper mapping.²⁴⁹ When a reference genome contains repetitive regions and/or structural variants, reads may map incorrectly and give rise to false-positive calls. Many variant-calling algorithms employ some degree of filtering, using either hardcoded or learned thresholds on metrics, such as base quality and strand bias, but personal experience suggests that even further filtering is typically necessary. It is here that a combination of both bioinformatics expertise and knowledge of the target genome is most helpful.

In our work on *Mtb*, we call SNVs using SAMtools mpileup, then remove all variants called within the ~10% of the *Mtb* genome that is highly repetitive; this is done by providing a list of the repetitive region coordinates in BED (browser extensible data) format to the BEDTools subtract program.²⁵⁰ The next step depends on the nature and scale of the genomic epidemiology investigation.

When examining a single outbreak or another set of closely related strains, our approach is to use a custom Python script to read in all of the variant positions in all of the VCF files (**Figure 1-5**) and keep only those that differ in at least one file. In this way, we are able to quickly discard the hundreds or thousands of SNVs that are identical across all sequenced outbreak isolates and are thus not informative for downstream transmission analysis. The script outputs a tab-delimited text file in which each row corresponds to a variant position; there are two columns for each sequenced isolate, one with the base call at that position (either the wild-type base appearing in the reference genome or a SNV) and one with the associated quality score, the scale of which depends on the variant caller used.

We next apply a distance filter, looking for any variant positions within 50 bp of another variant; in other words, we search for dense clusters of SNVs that are suggestive of mapping errors. We then apply a quality filter but in a flexible way. Many microbial genomics pipelines employ a strict quality threshold—if a variant's quality score does not meet or exceed the threshold, it is discarded. However, through frequent manual review of the data, it has often been found that these variants can indeed be trusted and ought not to be thrown out. Thus, it is recommended to

examine each row in the table, flagging those where at least one isolate achieves a minimum quality score (typically 200 when SAMtools mpileup is used) and discarding those where no isolate achieves the minimum score. We then manually inspect the remaining rows—usually these amount to no more than 20–100, depending on the scale of the outbreak—and assess each isolate’s base call and associated quality to arrive at a final data set of concatenated variation in FASTA format. While this manual inspection can often be done by examining data in the VCF file alone, it can sometimes be helpful to manually inspect the alignment using a BAM file viewer.

In larger studies retrospectively examining the many strains, clustered or otherwise, present in a region over time, or in the real-time prospective sequencing initiatives underway at national public health reference laboratories, this approach, which requires a certain degree of human intervention, is clearly not feasible. Instead, the bioinformatics pipelines supporting these studies use both repeat masking and carefully chosen thresholds on read depth, allele frequency, quality score, and strand bias to automatically filter out likely false positives and arrive at a final set of SNVs relative to the reference strain for each isolate. This approach may filter out some true-positive SNVs not meeting the various thresholds, but when the goal is simply surveillance and cluster identification rather than transmission inference, this automated approach is sufficient. Specific clusters can then be followed up in a second, less stringent analysis using an approach similar to that shown in **Figure 1-5**.

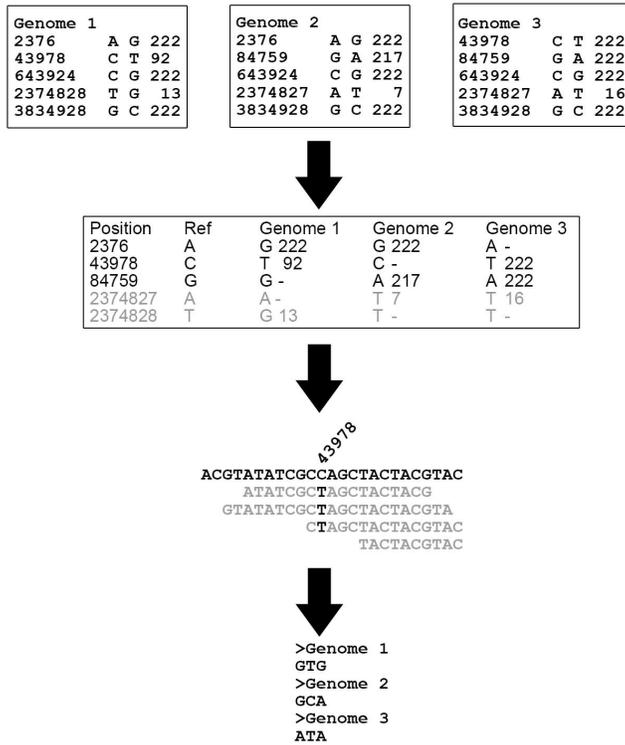


Figure 1-5. Identifying transmission-informative variation. Sequencing pathogen genomes from three cases yields three VCF files (top row). In this simplified example, the first column of each VCF contains the position, the second column contains the reference base, the third column contains the variant base call, and the fourth column contains the variant quality score. The variants called across all three isolates can be summarized in a matrix (second row). Here, the positions 643924 or 3834928 have not been included, as they are identical in all three genomes sequenced and are not informative for transmission—only bases that vary among the outbreak cases are included. In reviewing the quality scores and positions, the grayed-out rows would be excluded, positions 2374827 and 2374824, from subsequent analyses, as they have low-quality scores and are within 50 bp of each other, suggesting a mapping error. The third row indicates a manual review of the base called at positions 43978 in genome 1—although the variant was called with a low score of 92, manual inspection reveals that the low score is likely due to low coverage—only 3×—but that the variant call is true. The final row shows the concatenated variation for each genome in FASTA format.

Pipelines

Several bioinformatics pipelines have been developed to facilitate reference mapping and variant calling, combining the assembly and variant-calling steps into a single analysis; these might be organism specific, such as TGS-TB¹⁹⁶ or COMPASS-TB²¹⁴ for *Mtb* or wgsa.net for *S. aureus*,¹⁹⁷ or they might be organism agnostic, such as Snippy,²⁵¹ the PHENix²⁵² pipeline used by PHE, the CFSAN SNP²⁵³ pipeline used by the U.S. Food and Drug Administration, and SNVPhyl,²⁵⁴ used by PHAC. Multiple commercial solutions (reviewed by Wyres *et al.*²⁵⁵) are also available, including pipelines developed by the makers of several sequencers. Illumina's BaseSpace offers cloud-based analytics from both Illumina developers and external collaborators, while TorrentSuite offers analytics directly on the Ion family of sequencers. As genomics moves toward more routine clinical use, benchmarking these pipelines will become critically important. A set of best practices for evaluation was recently described by Olson *et al.*,²⁴⁹ while Budowle *et al.*²⁵⁶ describe the criteria for validating sequencing as a microbial forensics tool. Comparing a web-based tool like TGS-TB, targeted at research-oriented users with only a few genomes, to the command-line COMPASS-TB pipeline, designed to handle the real-time processing of every TB genome generated in UK reference laboratories, illustrates the features that make a pipeline scalable to a clinical level: the ability to handle hundreds of input sequences at a time on a high-performance computing back end; a well documented, transparent, and modular workflow; version control for both software tools and databases to promote reproducibility; the ability to integrate with other software platforms used in the clinical laboratory for sample tracking or reporting; and outputting a carefully designed report that meets regulatory standards for medical test reporting.

1.6.6 Step 4: rapid resistance prediction

Before discussing how transmission inference proceeds from assembled genomes, it is worth briefly describing how genomic data can be leveraged to predict an isolate's antibiotic susceptibilities. Appropriate antibiotic therapy is a cornerstone of TB treatment, but *Mtb*'s slow growth means that gold-standard, culture-based DST results take up to 8 weeks,²⁵⁷ during which time the patient may be on an incorrect treatment regimen. More rapid culture-based methods have been proposed, and molecular tests with turnaround times in hours, not days, are also

routinely used, including hybridization-based line-probe assays and real-time polymerase chain reaction (PCR).²⁵⁸ Concordance between molecular and phenotypic assays varies widely—while molecular methods for detecting resistance to INH and RIF are over 92% and 97% concordant with phenotypic results,²⁵⁹ molecular assessments of ethambutol and streptomycin resistance are closer to 50% concordant.²⁶⁰

For a molecular assay to be sensitive enough to replace culture-based DST, two conditions must be met. First, the suite of mutations conferring resistance to a particular drug must be well defined. For RIF, in which ~95% of resistance arises owing to mutations in a single gene (*rpoB*),²⁶¹ this is straightforward; for other drugs, genotype–phenotype correlations are not as robust.²⁶² This is particularly problematic for first-line therapies, such as pyrazinamide (PZA). Mutations in *pncA* constitute the primary mechanism of PZA resistance;²⁶³ however, phenotypically resistant isolates have been identified without *pncA* mutations. In this and many other discordant scenarios, other mutations are clearly involved, but many have yet to be revealed; thus, *pncA* remains the only target of value for predicting PZA resistance.²⁶⁴ However, the list of resistance-associated mutations in *Mtb* is growing thanks to large genomic studies^{265–268} and initiatives like ReSeqTB, a data-sharing platform developed through a unique collaboration among the WHO, CDC, the Stop TB New Diagnostics Working Group, the Critical Path Institute, FIND, and academic TB experts.²⁶⁹ As the number of sequenced *Mtb* genomes with linked metadata describing their phenotypic DST results grows, algorithms like that proposed by Walker *et al.*²⁶⁷ will be able to continuously self-update, identifying more and more resistance-associated mutations.

Second, for a molecular assay to replace culture-based DSTs, the assay itself must incorporate as many resistance-conferring mutations as possible. Tools like the INNO-LiPA Rif.TB test (Innogenetics NV, Belgium) and the GenoType MTBDRplus assay (Hain Lifescience GmbH, Germany) are limited to specific mutations in first-line drugs,^{270,271} with the recent GenoType MTBDRsl adding mutations in four genes associated with resistance to second-line antibiotics. Whole-genome sequencing offers a clear advantage here. Because it considers the entire genome, a single genomic analysis can scan for all known resistance mutations. This approach is at the

heart of recent online tools for predicting *Mtb* antibiotic susceptibilities directly from genomic data. KvarQ,²⁷² PhyResSE,²⁷³ and Mykrobe²⁷⁴ take FASTQ files as input and perform rapid, *in silico* phenotyping by scanning the short reads for a catalogue of resistance- and lineage associated mutations. Not having to assemble the reads into complete genomes means results are available within minutes, and the mutation catalogues can easily be updated as more resistance-conferring mutations are identified.

While resistance typing from genomic data is straightforward in *Mtb*, where resistance arises solely through point mutations, for other pathogens, resistance-conferring DNA is acquired from other sources,²⁷⁵ via transmissible plasmids, integrons, and transposons.²⁷⁶ Tools, such as SRST2²⁷⁷ and SEAR,²⁷⁸ take FASTQ files as input but must include a reference mapping step so as to capture resistance genes carried on MGEs. Accurate assembly of MGEs using short-read sequencing technology can be challenging, although hybrid assemblies incorporating long-read data are facilitating plasmid reconstruction.^{188,279} Predicting drug susceptibility from genomic data can facilitate rapid, targeted prescribing, improving patient outcomes and enhancing antimicrobial stewardship. However, it is not without its challenges. Mutations and gene presence/absence alone do not solely determine phenotypic resistance and the associated success or failure of antibiotic therapy—resistance may depend on gene expression levels,²⁸⁰ and treatment failure can be influenced by host genetic polymorphisms.²⁸¹ Nevertheless, for slow-growing organisms like *Mtb*, genomics offers a rapid alternative to traditional DSTs that is already affecting patient care.^{171,214}

1.6.7 Step 5: making the links

Returning to our reference-mapped genomes, we now have a FASTA file of the concatenated variation in each pathogen genome taken from our outbreak cases (**Figure 1-5**). Isolates that may have previously appeared identical via genotyping can now be distinguished from each other owing to the significantly improved resolution provided by genomics^{140,158,282,283}—not only can we now begin to assess which isolates are most closely related, potentially representing person-to-person transmission (**Figure 1-4**), but we can also distinguish between reinfection and relapse in diseases that may not always achieve cure, including *Clostridium difficile* and TB.^{119,120,284}

In reconstructing an outbreak from genomic data, it is easier to rule out transmission than it is to rule it in—research into HIV phylogenetics in the context of criminal prosecution makes this explicitly clear.²⁸⁵ In *Mtb*, Walker *et al.*¹⁴⁹ suggest that isolates >12 SNVs apart are not likely to be related epidemiologically. Nevertheless, drawing inferences around the precise route of transmission of closely related isolates—fewer than five SNVs apart for *Mtb*—is a useful exercise when genomics is used as part of a real-time outbreak-management strategy. In the earliest *Mtb* genomic epidemiology studies, clinical and epidemiological information, such as AFB smear results, site of infection, presence of cavitory disease, infectious period, named contacts, and named locations, was manually reviewed in the context of genomic data to establish plausible transmission events or clusters.^{74,149} More recently, automated methods for inferring transmission have been introduced that are extensible to other pathogens, including viruses. Tools like TransPhylo,¹⁴⁴ OutbreakTools,²⁸⁶ SEEDY,²⁸⁷ and within-host coalescent analysis in BEAST¹⁵⁴ take, as input, concatenated genomic variants from outbreak isolates or a phylogenetic tree showing this variation. These methods combine complex epidemiological and evolutionary models to infer the transmission tree—the chain of person-to-person infection events—that best explains the observed genomic variation; some methods are also able to assign a date interval within which an infection event likely occurred. Several tools also account for within-host genetic diversity, which can complicate manual transmission inference when dealing with pathogens that have latent periods or periods of asymptomatic carriage, including *Mtb* and *S. aureus* infection (**Figure 1-4**).^{162,288}

In our work, we use a hybrid approach to reconstructing transmission. We first use TransPhylo to infer a putative transmission network from our concatenated genomic variation alone.¹⁴⁴ This initial network must then be manually refined using our clinical and epidemiological data. This refinement is necessary for two reasons. First, the initial network contains many more edges—potential person-to-person transmission events—than would have actually occurred in the outbreak. In other words, for any one person in the network, the genomic data might suggest multiple different sources of that person’s infection. Edges in automatically inferred networks are also often bidirectional, indicating a potential infection event between persons A and B but not the direction of that event. Second, the evolutionary models underlying automated-

transmission inference tools are not well suited to capturing the variable tempo of mutation in organisms like *Mtb*.¹⁴⁷ The development of TransPhylo demonstrated that transmissions strongly supported by epidemiology may not be captured in the inferred network because of the model's parameters.¹⁴⁴

In refining the network, we rely on the available clinical and epidemiological data associated with each case. We attempt to prune the network, removing edges that we deem unlikely to represent true transmission events between individuals. Using clinical data, we remove edges emanating from individuals with non-infectious forms of TB, for example non-pulmonary forms of the disease with no evidence of AFB in a patient's sputum. We also remove edges that are unlikely to represent transmission given what we know about the location and timing of cases. In outbreaks spanning geographically distinct locales, we can eliminate edges between cases without reported travel histories. Similarly, we can use an individual's prior TST or chest X-ray findings to narrow down transmission possibilities—repeated screening in a large outbreak will often reveal several individuals with a documented window of TST conversion; individuals who had initiated treatment and were therefore removed from the pool of potential infectors before this window can be ruled out as a source of these cases' infection. We can also confirm edges based on named contact or location data—if an individual has multiple potential sources in the network but only one of those is a named contact or someone known to frequent the same locations, we select that as the most probable edge.

The nature of the clinical and epidemiological data that can influence a genomic epidemiology reconstruction will vary from pathogen to pathogen. In outbreaks of hospital-acquired infections, one must consider everything from patient movement between beds and wards to whether specific pieces of equipment were used in an individual's care,^{202,289} while in outbreaks of foodborne illness, data collection may have to span international or even intercontinental borders.^{203,290} Accurately reconstructing an outbreak from genomic data therefore requires thorough knowledge of the pathogen in question and the epidemiology of the resulting disease and a recognition that undertaking such an investigation may have legal, economic, and/or privacy ramifications, as reviewed by Gilchrist *et al.*¹⁵¹ It is important to note that, just as

epidemiological data might influence the decision to use genomics, so can genomics inform the decision to undertake an epidemiological investigation. For example, investigating a nosocomial outbreak of an antibiotic-resistant pathogen can be time consuming and expensive. Genomics can quickly establish whether suspected outbreak isolates are related and an investigation is warranted or whether the putative outbreak represents a series of unrelated introductions into the hospital environment with no onward transmission.

1.6.8 Concluding thoughts

Genomics has had a profound impact on our ability to understand infectious disease epidemiology. When coupled with active surveillance programs, genome sequencing is solving more outbreaks,²⁹¹ suggesting new modes of transmission,²⁹² and revealing new reservoirs of disease.²⁰⁴ Furthermore, the routine, real-time use of genomics within the clinical reference laboratory means that genomics can do more than identify epidemiological links between cases—it can also be used as a tool for the rapid diagnosis and resistance typing of an isolate, as has recently been demonstrated for *Mtb*.²¹⁴ And, when the lens of genomics is turned upon a larger set of samples collected across time and space, we gain a deeper understanding of disease ecology. Genomic investigations of *Shigella sonnei*²⁹³—an emerging cause of dysentery in the developing world—and *Vibrio cholerae*¹⁹⁸ have revealed the role that population expansion plays in a disease’s success. Sequencing of *S. pneumoniae* has shown how capsular switching mediates vaccine escape,²⁹⁴ while work on *Neisseria gonorrhoeae* charts the dispersal of cefixime resistance.²⁹⁵ As the number of publicly available genome sequences increases, so does our ability to mine sequence data and carry out comparative genomic analyses, leading to vaccine development, new drug targets, improved diagnostics, and the identification of novel drug resistance and virulence loci.²⁹⁶

Despite the promise of genomics,²⁹⁷ the community is currently dealing with several challenges, such as validating genomics against existing molecular tools, accrediting ever-changing bioinformatics pipelines for use in a clinical environment, and communicating complex genomic information to end users. Moving forward, training the next generation of disease detectives represents the next great hurdle—we need analysts with graduate-level skills in bioinformatics,

evolutionary biology, and infectious disease microbiology and epidemiology. Challenges aside, the benefits of whole-genome sequencing in a public health environment are clear, and all are looking forward to the exciting and transformative insights yet to come as groups around the world harness the power of sequencing.

1.7 Accelerating TB Elimination in Low-incidence Settings: the Role of Genomics

In a recent ERJ article, Lönnroth *et al.* proposed a framework to accelerate progress towards tuberculosis (TB) elimination in low-incidence settings.¹³⁵ In it, they outline eight priority areas and multiple interventions that align with the World Health Organization’s post-2015 global TB strategy.²⁹⁸ This framework is to be applauded and the recognition that elimination in low-incidence countries is a unique problem, where infection occurs amongst the most difficult-to-reach individuals. Although “new research and tools” is one of the framework’s areas, it overlooks an important new technology that is changing our understanding of TB and our approaches to diagnosis, phenotyping, and treatment—whole genome sequencing (WGS) of *Mycobacterium tuberculosis* (*Mtb*) isolates from cases of active TB.^{299,300}

In contrast to genotyping, which interrogates ~0.5% of the *Mtb* genome, WGS reads the entire 4.4 Mbp of sequence—with current sequencing technologies this can be done in under a day at a cost of ~\$50–\$100/genome depending on the laboratory.³⁰¹ Many federal public health agencies have invested substantially in genomics and are using it routinely in medical microbiology,³⁰² with TB representing the ideal use case. The *Mtb* genome is uncomplicated—the global population of *Mtb* is clonal and the genome comprises a single chromosome in which variations arise through point mutations—facilitating downstream bioinformatics analyses such as prediction of antimicrobial resistance phenotypes. Additionally, most TB diagnostic work is performed in well-equipped reference laboratories, where centralization permits integrating WGS data with the data streams necessary for diagnosis, surveillance, outbreak management, and research, and where the need for accredited operating procedures is accelerating the standardization of TB genomics protocols—Public Health England is already using WGS routinely alongside its traditional mycobacteriology laboratory pipeline in its Birmingham Public Health laboratory.³⁰²

There are many areas of the framework in which WGS is poised to support efforts in tuberculosis elimination. First, WGS is able to resolve TB transmission dynamics, including individual transmission events, to a degree not possible with traditional genotyping. This genomic epidemiology approach has been used to reconstruct single transmission events, local outbreaks, and regional epidemics,³⁰⁰ and is providing insights into both patterns of spread and characteristics of transmitters. Genomic epidemiology speaks to several priority action areas. Unlike genotyping, which simply clusters active cases, WGS can elucidate the order and direction of transmission, revealing common trends in TB outbreaks and identifying those individuals who are transmitting disease and those who aren't. This directly informs the framework's goals of describing local patterns of transmission amongst vulnerable populations and identifying individuals most at risk for transmitting disease (areas 2 and 3); it also enables us to prioritize contacts of these key transmitters for screening (area 4). Additionally, real-time WGS of specimens from new TB cases provides a core indicator of ongoing transmission for surveillance (area 6).

Another priority area is the prevention and care of drug-resistant TB (area 5), for which phenotypic methods are the current gold standard. These methods are culture-dependent and have turnaround times (TATs) of weeks to months; even rapid molecular methods with TATs of hours—such as the Cepheid GeneXpert and line probe assays—are limited to detecting only a handful of known resistance mutations. As much as 1/3 of isoniazid resistance cannot be explained by these canonical mutations,³⁰³ and no commercially available test can probe resistance to all anti-tuberculous drugs. WGS, in contrast, interrogates the entire genome, identifying resistance-associated mutations in hours to days³⁰⁴ and optimizing the prescription of appropriate treatment and supporting rational drug use (area 1). Although this requires a comprehensive database of well-described resistance-associated mutations, several such efforts are underway and more mutations are likely to be described as ever-increasing numbers of isolates are sequenced.

There are other, less obvious, benefits to incorporating WGS into the routine diagnosis, management, and surveillance of TB in low-incidence settings. Chief amongst these is that just

as early *Mtb* genomics efforts led to new molecular diagnostic tools based on a handful of sequenced genomes,²⁹⁹ the generation and sharing of thousands of *Mtb* genomes will likely lead to new molecular tools for use in higher-incidence settings. Genomics is stimulating TB research and reinvigorating a community that has experienced de-prioritisation (area 1), potentially leading to renewed political interest in TB programs and the specialized training and creation of central repositories. In addition, collaborative efforts—necessary to deploy WGS in the clinical laboratory, will act to build capacity at national and international levels, which is essential for the elimination of TB in low- and high-incidence settings.

In conclusion, genomics stands to significantly enhance TB elimination efforts through direct and indirect routes. When combined with the framework’s recommended interventions, it is believed that WGS has the potential to accelerate progress towards TB elimination in low-incidence countries, with the knowledge gained in these settings working to support the final priority action area—informing TB prevention, care, and management in countries with a high burden of disease.

Chapter 2: Universal Genotyping for Tuberculosis Prevention Programs: A Five-Year Comparison with On-Request Genotyping

2.1 Background

Despite declining case rates, tuberculosis (TB) remains a public health issue in Canada and other low-incidence countries.¹³⁵ Here, a substantial proportion of TB diagnoses occur in persons born outside Canada and represent reactivation of latent TB infection.^{6,135} However, outbreaks and endemically circulating strains also contribute to incidence rates.^{74,166,305} Interruption of these transmission chains requires an understanding of regional epidemiology. Techniques such as 24-locus mycobacterial interspersed repetitive unit–variable-number tandem repeat (MIRU-VNTR) genotyping can provide valuable insights into the potential extent of local TB transmission by using clustering as a proxy; thus, many low-incidence settings have incorporated MIRU-VNTR genotyping into standard practice.^{121,306,307}

Several laboratories now perform universal genotyping,^{307–312} in which all culture-positive isolates from a region are prospectively genotyped by one or more molecular methods. While published reports have examined clustering rates and other metrics related to universal genotyping programs,^{313–315} there are no reports directly comparing the results of universal genotyping to those of an on-request genotyping program over the same time period in the same region.

In the Province of British Columbia (BC), Canada, *Mycobacterium tuberculosis* isolates are MIRU-VNTR genotyped by the BC Centre for Disease Control (BCCDC) Public Health Laboratory (BCPHL). From 2009 to 2014, genotyping was done only when requested by BCCDC TB Services personnel. However, a recent province-wide retrospective molecular epidemiology research study later genotyped all culture-positive BC isolates from 2005 to 2014 ($n = 2,290$) to describe the complete genotypic landscape of TB in BC, the results of which are detailed in **Chapter 3**. This data set was used to compare the insights derived from the on-

request genotyping performed between 2009 and 2013 to those later revealed through genotyping of all of the remaining isolates during this period. Given the significant costs, time, and effort associated with the implementation of universal genotyping, it was important to assess the epidemiological value it adds in a low-incidence setting such as BC, where >75% of TB cases occur in persons born outside Canada and are likely not due to local transmission.^{6,9}

2.2 Materials and Methods

2.2.1 On-request genotyping data

The BCPHL performs routine TB diagnostics, phenotypic drug susceptibility testing, and 24-locus MIRU-VNTR genotyping for all culture-confirmed cases in BC. Until mid-2014, MIRU-VNTR genotyping was performed only when requested by a clinician—typically to support outbreak investigations and contact tracing efforts—with all requests recorded in a spreadsheet. This spreadsheet was used to identify all of the genotyping requests received between 1 January 2009 and 31 December 2013—the last full calendar year before universal genotyping was implemented. On the basis of the information contained in the spreadsheet, the reason for each request was coded as (i) suspected possible transmission, (ii) distinguishing relapse from reinfection, or (iii) suspected false-positive results. For inquiries regarding possible transmission, it was noted whether the request asked for comparison to a specific individual(s), to a known outbreak, or to the general database.

2.2.2 Universal genotyping data

In **Chapter 3**, a retrospective genotyping analysis of culture-positive *M. tuberculosis* isolates diagnosed in BC between 2005 and 2014 will be described; here, the subset of isolates received between 2009 and 2013 ($n = 1,136$) and an additional 39 isolates requested for genotyping during this period but from specimens received prior to 2009 were examined. For case-based analyses, the study sample excluded false-positive TB diagnoses ($n = 3$) and the second record of a reoccurrence, leaving a total of 1,158 cases. Briefly, *M. tuberculosis sensu stricto* isolates were genotyped by standard 24-locus MIRU-VNTR methods⁷⁷ and linked to individual-level clinical,

demographic, and epidemiological data extracted from the BCCDC Integrated Public Health Information System (iPHIS). Postal codes were used to obtain the corresponding census dissemination area (DA) for each case, which was then linked to the 2006 Canadian Marginalization Index (CAN-Marg) to determine the deprivation index quintile.³¹⁶

2.2.3 Statistical analysis

Data were analyzed and presented as means with standard deviations and relative frequencies, as appropriate. Logistic regression was used to estimate the odds ratio (OR) and 95% confidence interval (CI) for the association between genotype requested to confirm/refute transmission (yes/no) and MIRU-VNTR genotyping clustered (yes/no). A cluster was defined as ≥ 2 isolates with identical 24-locus MIRU-VNTR genotyping patterns by using a stringent perfect type match, and each cluster was labeled with a unique identifier (MClustID). To obtain the adjusted OR (aOR), variables were selected *a priori*, which included age group, gender, birthplace (Canada/outside Canada), and the presence of one or more risk factors (HIV, illicit drug use, or alcohol misuse) known to be associated with TB transmission and therefore genotype clustering.³¹⁷ Only cases with complete data for all variables were included in the model ($n = 910$). A secondary analysis was conducted on a subset of the 2009 to 2013 data (2013 quarter 3 [Q3] and Q4 excluded) to explore the possibility that the relationship between genotypic clustering and request status was influenced by the large increase in requests during the last two quarters of 2013. An additional analysis to examine risk factors in relation to genotype requests and clustering status used case records with complete risk factor data ($n = 916$). Characteristics of all clusters with ≥ 3 persons (i.e., growing clusters) were analyzed, and the predominant birthplace was assigned as Canada where $\geq 50\%$ of the persons in the cluster were born in Canada; otherwise, the predominant birthplace was categorized as outside Canada. The cluster growth rate was calculated as the average increase in case counts per quarter over the study period, and linear regression was used to test the relationship of growth rate, cluster size, and birthplace on cluster proportion requested. All analyses were executed in R (v3.3.1).

2.3 Results

2.3.1 The genotype request proportion was smaller than the genotypic clustering proportion

The study sample included 1,175 isolates, consisting of 1,136 culture-positive *M. tuberculosis* specimens received by the BCPHL from 2009 through 2013 and 39 isolates requested during the study period that were received prior to 2009 (**Figure 2-1**). During this time, clinicians submitted 194 genotyping requests involving 309 isolates from 296 individuals, including 13 isolates from TB recurrences. The quarterly request proportion varied over time, averaging 20.5% before 2013 Q3, at which point requests increased (**Figure 2-2**). Of the 1,136 specimens received by the BCPHL during the study period, only 271 (23.8%) had genotyping requested specifically to confirm or refute suspected transmission (**Table 2-1**) However, the subsequent universal genotyping analysis revealed an overall provincial genotypic clustering proportion of 38.0%, meaning that prior to universal genotyping, on-request genotyping captured fewer clusters.

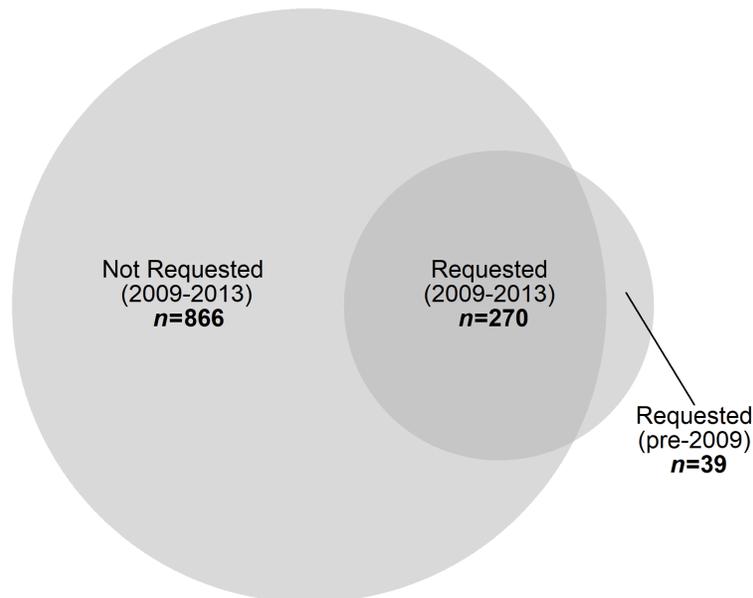


Figure 2-1. Study sample request status. Diagram representing the request status of the study sample ($n = 1,175$) for isolates with genotyping requested (all reasons) from 2009 through 2013, which included all genotyped isolates from specimens received at the British Columbia Centre for Disease Control Public Health Laboratory from 2009 through 2013, and those requested for genotyping with a specimen received date prior to 2009.

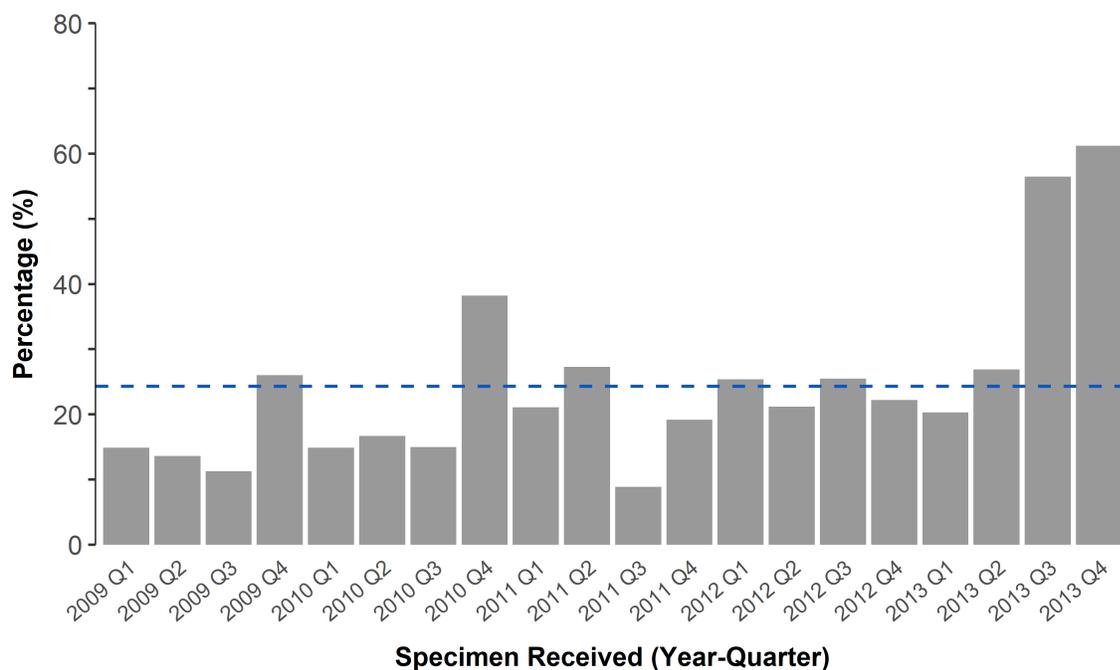


Figure 2-2. Quarterly genotype requests. Percentage of isolates requested for genotyping (all reasons) displayed by year and quarter specimens were received at the British Columbia Centre for Disease Control Public Health Laboratory from 2009 through 2013. Dashed line represents the mean percentage of genotyped isolates requested per quarter.

Table 2-1. Genotype request reasons. Frequency of genotype requests by reason, British Columbia, 2009–2013 ($n = 300$)^a.

Request Reason	n (%) ^b
Transmission	
Specified case comparison	41 (13.7)
Specified outbreak comparison	111 (37.0)
General database comparison	119 (39.7)
Relapse or reinfection	12 (4.0)
Specimen mix-up/cross-contamination	17 (5.7)

^aIncluded are all cases who were subjects of genotyping requests ($n = 296$). Four individuals were the subjects of multiple genotyping requests for different reasons; here, each request is counted separately ($n = 4$).

^bPercentages have been rounded and may not add up to 100%.

2.3.2 Requests reflected suspected community transmission and known risk factors

Most requests (90.3%) were made during contact investigations to confirm or refute transmission, although few named specific individuals (**Table 2-1**). Instead, most requests asked for a comparison against a specific outbreak genotype or the general database. When a specific comparator was identified ($n = 152$ requests)—either an individual or a specific outbreak genotype—a match between the requested strain and comparator was observed in 83 instances (54.6%). Upon examination of all isolates requested to determine possible transmission, it was found that 67.5% (183/271) matched at least one other isolate by MIRU-VNTR genotyping during the study period. Requests to differentiate relapse from reinfection ($n = 12$) and requests to investigate potential laboratory errors ($n = 17$) were less frequent.

Next, the characteristics of individuals for whom genotyping was requested to confirm or refute transmission ($n = 269$ after the exclusion of two individuals whose genotype was requested on more than one occasion to investigate transmission) versus all other cases in the study sample representing true positive TB diagnoses were compared (**Table 2-2**). It was found that proportionally more requests were made for individuals in the 35- to 54-year age group, males, those born in Canada, and persons with one or more risk factors (HIV, illicit drug use, or alcohol misuse).

Table 2-2. Study sample characteristics. Demographic characteristics of the study sample^a ($n = 1,158$), comparing individuals whose isolates were requested for genotyping to confirm/refute transmission ($n = 269$) versus all other samples ($n = 889$).

Characteristic	Genotyping Requested to Confirm/Refute Transmission		<i>p</i> -value ^b
	Yes <i>n</i> (%)	No <i>n</i> (%)	
Age, years			
0–34	60 (23.6)	194 (76.4)	<0.001
35–54	111 (32.5)	231 (67.5)	
55–74	66 (21.5)	241 (78.5)	
75+	32 (12.5)	223 (87.5)	
Gender			
Male	168 (24.7)	513 (75.3)	0.188
Female	101 (21.2)	376 (78.8)	
Birthplace ^c			
Canada	158 (51.6)	148 (48.4)	<0.001
Outside Canada	105 (12.9)	709 (87.1)	
Risk Factors ^d			
None	131 (16.6)	657 (83.4)	<0.001
≥1	70 (54.7)	58 (45.3)	

^aExcluded false-positive TB diagnoses ($n = 3$) and counted each individual once by excluding the second record from reoccurrences ($n = 14$).

^bChi-square test.

^cData unavailable ($n = 38$).

^dThe risk factors are HIV, illicit drug use, and alcohol misuse. Data unavailable for one or more risk factors, $n = 242$.

2.3.3 Universal genotyping improves cluster identification

Province-wide, retrospective universal genotyping revealed how many clusters and how many clustered individuals were missed during the on-request period. From 2009 through 2013, 94 genotypic clusters were observed in BC, ranging in size from 2 to 53 cases (mean = 5) and involving a total of 432 individuals. On-request genotyping missed 54 (57.4%) of these clusters and 130 (30.1%) clustered individuals (**Table 2-3**).

Ten clusters (10.6%), with an average of three isolates per cluster, were fully identified through on-request genotyping; 30 clusters (31.9%) were partially identified (**Table 2-3**; **Figure 2-3**). These partial clusters tended to be larger (9.1 ± 10.7 persons/cluster) than those that were either missed or fully identified (≤ 6 persons/cluster). The mean proportion of requested cases within a partially identified cluster was 40.5%. Clusters described as predominantly Canadian-born ($n = 29$) were more likely to be partially or fully requested (**Table 2-3**).

Table 2-3. Characteristics of MIRU-VNTR clusters. Characteristics of MIRU-VNTR clusters identified through universal genotyping categorized by the proportion of each cluster (none, partial or all) requested for genotyping to confirm or refute potential transmission.

Cluster Requested Proportion	No. of Clusters	Predominantly Canadian-born n (%)	Cluster Size Range	Mean Cluster Size ($\pm SD$)
None (0%)	54	10 (18.5)	2–6	2.4 ± 0.8
Partial (1–99%)	30	14 (46.7)	2–53	9.1 ± 10.7
All (100%)	10	5 (50.0)	2–5	3.0 ± 1.2

Abbreviation: *SD*, standard deviation.

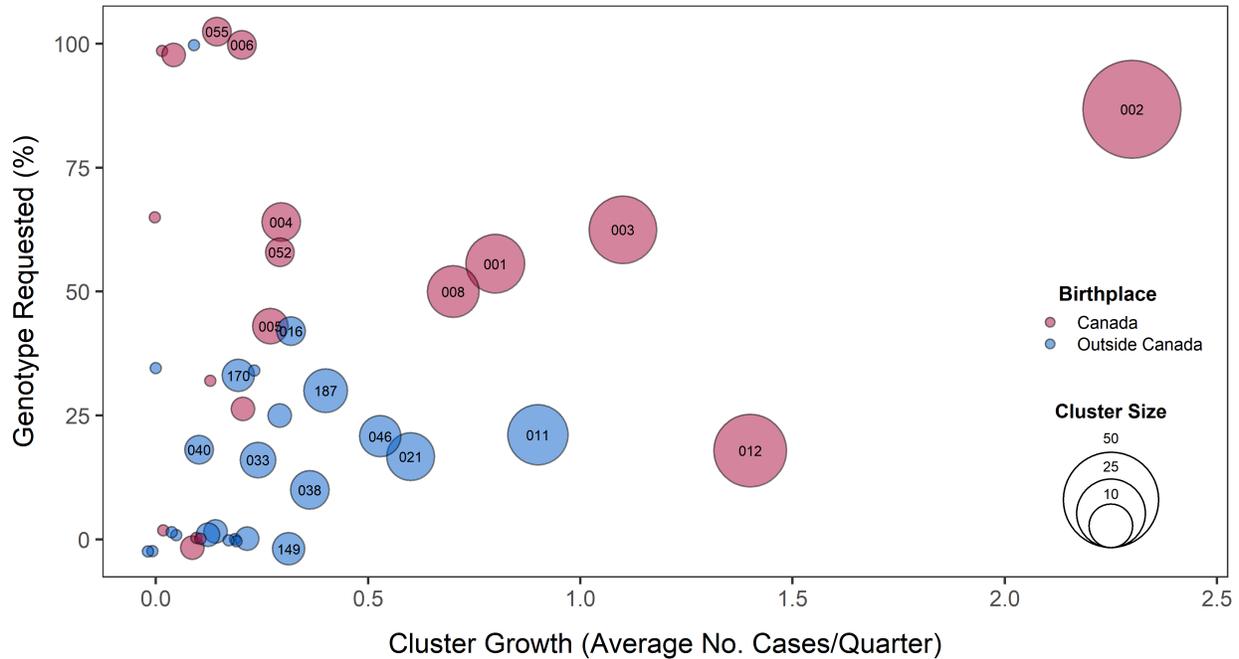


Figure 2-3. Proportion of each cluster requested by cluster growth over time. Bubble plot of the proportion of each cluster requested for genotyping to confirm or refute transmission, with the average cluster growth per quarter in BC from 2009 to 2013. Growing clusters had a minimum of three persons in the cluster over the study period. Bubbles are colored to indicate the predominant birthplace ($\geq 50\%$) of the individuals in each cluster and sized to represent the total number of genotypically clustered cases. Cluster identifiers are indicated for clusters with five or more cases.

Logistic regression analysis was used to examine the characteristics of those in genotypic clusters and found that individuals in the 35- to 54-year age group, males, those born in Canada, and persons with one or more risk factors (HIV, illicit drug use, or alcohol misuse) were more likely to belong to a cluster than to have a unique genotype (**Table 2-4**). It was observed that isolates that had a historical genotype request had greater odds of belonging to a genotypic cluster (aOR 2.3, 95%CI: 1.5–3.3); this effect size increased (aOR 3.3, 95%CI: 2.0–5.4) when the last two quarters of 2013 from the analysis were excluded (**Table 2-5**). Risk factors in relation to genotype requests and clustering status were examined and it was found that 258 (72.5%) of the 356 persons with clustered isolates had no risk factors identified (**Table 2-6**).

Table 2-4. Logistic regression. Analysis for the relationship between MIRU-VNTR genotypic clustering, as revealed by universal genotyping, and whether an isolate had originally been requested for genotyping to confirm or refute transmission.

Characteristic	Clustered ^a vs. Unique	
	Unadjusted OR (95%CI)	Adjusted OR (95%CI)
Requested		
Yes	4.6 (3.3–6.5)	2.3 (1.5–3.3)
No	Reference	Reference
Age, years		
0–34	Reference	Reference
35–54	1.7 (1.2–2.5)	1.5 (1.0–2.3)
55–74	0.9 (0.6–1.4)	1.0 (0.6–1.5)
75+	0.5 (0.3–0.8)	0.8 (0.5–1.3)
Gender		
Male	1.3 (1.0–1.7)	1.1 (0.8–1.5)
Female	Reference	Reference
Birthplace		
Canada	8.8 (6.2–12.3)	5.3 (3.5–7.8)
Outside Canada	Reference	Reference
Risk Factors ^b		
None	Reference	Reference
≥1	6.6 (4.2–10.2)	1.8 (1.0–3.0)

Abbreviations: CI, confidence interval; OR, odds ratio.

^aCluster: ≥ 2 isolates that share an identical genotype (24-locus MIRU-VNTR).

^bRisk Factors = HIV positive, illicit drug use, or alcohol misuse.

Table 2-5. Logistic regression with a restricted dataset. Logistic regression analysis for the relationship between MIRU-VNTR genotypic clustering and genotyping requested (2009–2013Q2) to confirm or refute transmission ($n = 813$), British Columbia.

Characteristic	Clustered ^a vs. Unique	
	Unadjusted OR (95%CI)	Adjusted OR (95%CI)
Requested		
Yes	8.5 (5.5–13.1)	3.3 (2.0–5.4)
No	Reference	Reference
Age, years		
0–34	Reference	Reference
35–54	1.8 (1.3–2.7)	1.6 (1.0–2.5)
55–74	1.0 (0.6–1.4)	1.0 (0.6–1.6)
75+	0.5 (0.3–0.8)	0.8 (0.4–1.3)
Gender		
Male	1.2 (0.9–1.6)	1.0 (0.7–1.4)
Female	Reference	Reference
Birthplace		
Canada	9.1 (6.3–13.0)	4.9 (3.2–7.3)
Outside Canada	Reference	Reference
Risk Factors ^b		
None	Reference	Reference
≥1	6.8 (4.3–10.8)	1.8 (1.0–3.1)

Abbreviations: CI, confidence interval; OR, odds ratio.

^aCluster: ≥ 2 isolates that share an identical genotype (24-locus MIRU-VNTR).

^bRisk Factors = HIV positive, illicit drug use, or alcohol misuse.

Table 2-6. Request status, risk factor, and clustering.
Relationship between genotype request status, risk factors^a and genotypic clustering (Yes/No), British Columbia, 2009–2013.

Characteristic	No. Isolates (%)	Clustered (<i>n</i>)	
		Yes	No
Requested			
<i>No Risk Factors</i>	131 (65.2)	73	58
<i>≥1 Risk Factors</i>	70 (34.8)	62	8
Not Requested			
<i>No Risk Factors</i>	657 (91.9)	185	472
<i>≥1 Risk Factors</i>	58 (8.1)	36	22

^aRisk Factors = HIV, illicit drug use, or alcohol misuse; data unavailable for 1 or more risk factor (*n* = 242).

2.3.4 Growing clusters were variably identified by on-request genotyping

To examine growing clusters, the data set was pruned to include only the 43 clusters with three or more persons and examined the cluster growth rate and the proportion of requested cases (**Figure 2-3** and **Figure 2-4**). Although request rates varied, Canadian-born clusters with higher growth rates were larger and tended to have proportionally more isolates requested for genotyping ($p = 0.003$). MClust-002, a previously described TB outbreak in BC,¹⁷³ was the largest cluster observed during the study period ($n = 53$) and had the highest average rate of growth (2.3 cases/quarter) and the largest number of clustered cases observed in a single quarter ($n = 9$). Within this cluster, an additional seven cases were identified through universal genotyping—six of these were early in the outbreak (2009 Q1). Two other recognized outbreaks, one previously described⁷⁴ (growth rate = 0.8 case/quarter) and the other spanning a more remote part of the province (1.1 cases/quarter), had partially requested isolates (44.4% and 37.5% of cases missed, respectively). MClust-012 involved an urban population with a high material deprivation index (**Table 2-7**). Here, only five of 28 individuals in the cluster had a genotyping request (**Figure 2-4**; **Table 2-7**), three of which were late in the outbreak (2013), and the requests for a 2009 and a 2010 isolate asked for comparisons to outbreak strains other than MClust-012. Requests were less common among clusters involving largely non-Canadian-born individuals, where the request rate in the three largest clusters (≥ 10 individuals) averaged 22.6% (**Table 2-7**).

Table 2-7. Genotype cluster characteristics. Characteristics of 24-locus MIRU–VNTR clusters comprised of ≥ 5 individuals, displayed as clusters that were predominantly Canadian- or non-Canadian-born, British Columbia, 2009–2013.

Cluster ID	Cluster Size (% Requested)	Predominant Birthplace ^a (%)	Median Age (IQR) years	Gender M:F	Predominant Community Type (%)	Risk Factors ^b (%)	Median Deprivation ^c Quintile
Canadian-born							
MClust-002	53 (86.8)	Canada (86.0)	51 (44–58)	12.2	Metro (71.7)	39.6	4.0
MClust-012	28 (17.9)	Canada (89.3)	46 (41–51)	1.8	Metro (89.3)	35.7	4.0
MClust-003	24 (62.5)	Canada (95.7)	44 (28–52)	2.4	Urban/Rural (41.7/41.7)	45.8	4.0
MClust-001	18 (55.6)	Canada (100.0)	46 (37–52)	0.8	Rural (50.0)	44.4	4.0
MClust-008	14 (50.0)	Canada (85.7)	51 (37–63)	1.3	Metro (85.7)	42.9	3.0
MClust-004	8 (62.5)	Canada (100.0)	47 (40–57)	1.0	Urban (75.0)	37.5	4.0
MClust-005	7 (42.9)	Canada (85.7)	57 (46–60)	2.5	Metro (85.7)	14.3	2.0
MClust-006	5 (100.0)	Canada (100.0)	25 (23–37)	0.7	Rural (80.0)	0.0	3.0
MClust-052	5 (60.0)	Canada (100.0)	53 (43–61)	NA ^d	Metro/Urban (40.0/40.0)	100.0	5.0
MClust-055	5 (100.0)	Canada (100.0)	42 (34–48)	0.7	Urban (80.0)	20.0	3.0
Non-Canadian-born							
MClust-011	19 (21.1)	Philippines (94.4)	42 (33–53)	2.8	Metro (68.4)	0.0	4.0
MClust-021	12 (16.7)	Philippines (100.0)	44 (29–51)	0.7	Metro (83.3)	0.0	3.5
MClust-187	10 (30.0)	<i>Mixed</i> (88.8)	80 (55–88)	0.7	Metro (100.0)	0.0	3.0
MClust-046	9 (22.2)	India (77.8)	70 (56–75)	1.2	Metro (88.9)	0.0	2.0
MClust-038	8 (12.5)	China/Hong Kong (100.0)	79 (72–81)	3.0	Metro (100.0)	0.0	3.5
MClust-033	7 (14.3)	<i>Mixed</i> (85.8)	68 (42–76)	0.8	Metro (100.0)	0.0	3.0
MClust-149	6 (0.0)	India (83.3)	50 (34–76)	0.5	Metro/Urban (50.0/50.0)	0.0	2.5
MClust-170	6 (33.3)	China (83.3)	68 (55–75)	1.0	Metro (100.0)	0.0	3.0
MClust-016	5 (40.0)	Philippines (100.0)	39 (32–41)	0.7	Metro/Urban (40.0/40.0)	0.0	3.0
MClust-040	5 (20.0)	<i>Mixed</i> (80.0)	75 (35–86)	0.7	Metro (100.0)	0.0	3.0

Abbreviation: IQR, interquartile range.

^aInformation for birthplace was unknown for six persons; percentage represents those with complete data. *Mixed* indicates various Asian countries.

^bOne or more risk factors for transmission (HIV, illicit drug use, or alcohol misuse); data unavailable ($n = 50$); percentage represents those with complete data.

^cCanadian Marginalization Index,³¹⁶ material deprivation (quintile 1: least deprived, quintile 5: most deprived); data unavailable ($n = 23$).

^dAll individuals were male.

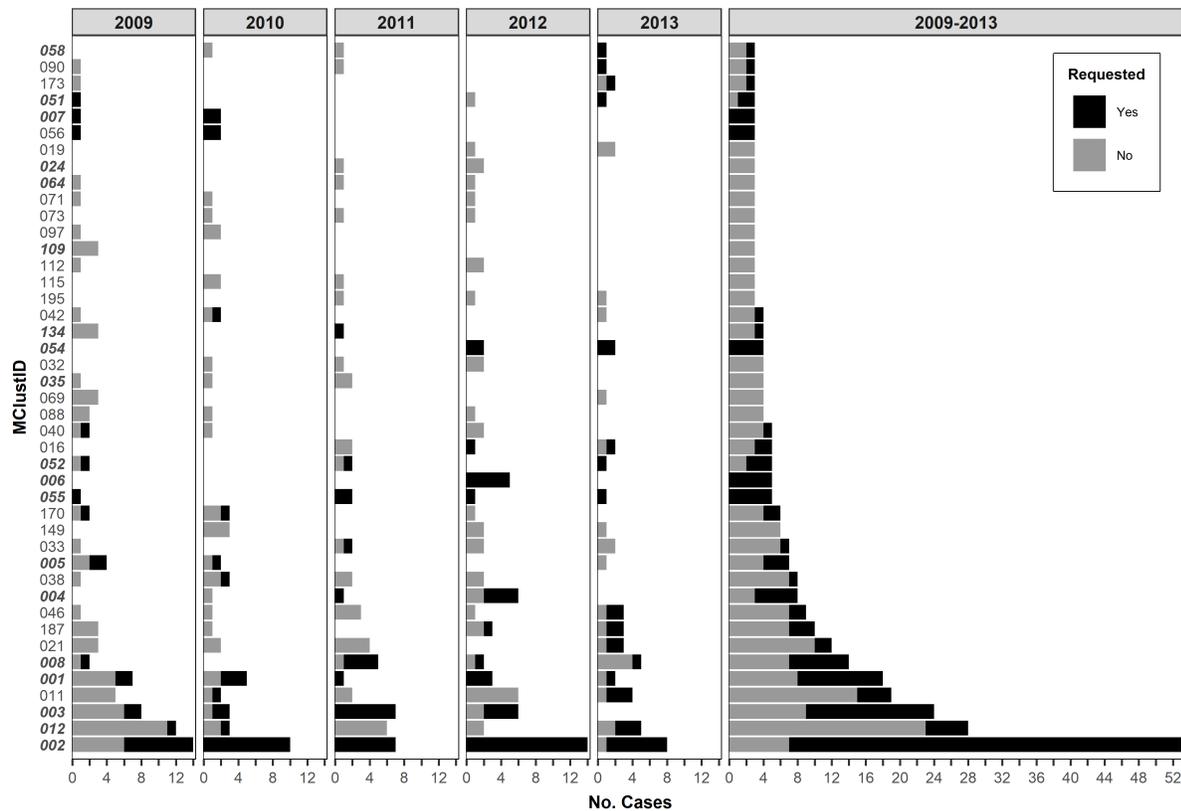


Figure 2-4. Cluster growth by genotype request status. Annual cluster growth and overall cluster size for all clusters with three or more persons in BC from 2009 to 2013. Bars are colored by genotype requested (yes/no). Twenty-four-locus MIRU-VNTR genotyping cluster identifiers (MClustID) in bold italics represent clusters that are composed of predominantly Canadian-born persons.

2.4 Discussion

In low TB incidence settings, clinical laboratories considering universal genotyping must demonstrate that it offers substantial epidemiological insights beyond those from an on-request service. This study leveraged a unique situation, in which five years of an on-request genotyping program was compared to the information later gained from retrospective genotyping of all of the isolates during this period, to generate the evidence to justify ongoing universal genotyping.

During the on-request period, the existence of many genotypic clusters and the full extent of many other clusters were missed. MIRU-VNTR genotyping overestimates the clustering of related isolates, particularly for clusters involving non-Euro-American *M. tuberculosis*

lineages.³¹⁸ With 62% of BC's cases attributable to non-Euro-American lineages (see **Chapter 3**), some of the missed clusters are likely pseudoclusters and do not reflect true local transmission. However, clusters involving Canadian-born persons that do likely represent local transmission were also partially or fully missed by on-request genotyping. Whole-genome sequencing (WGS) of all clustered isolates was later undertaken to provide a more accurate quantification of local transmission within BC, as well as strain-specific insights into drug resistance and transmissibility (**Chapter 8**).

Genotyping requests were most often used to investigate suspected community transmission, particularly in individuals with known risk factors. MIRU-VNTR genotyping results confirmed many potential transmission events, but specific suspicions, in which an individual or outbreak strain comparator was noted in the request, were less frequently correct. This suggests that clinicians understood the risk factors for transmission but that the underlying epidemiological networks were not as clear. Universal genotyping provides a bias-free method to identify connections between cases and reveal possible transmissions between individuals who do not fit traditional risk profiles.

In a secondary analysis, restricting the data to include only dates prior to the spike in requests (2013 Q3 and Q4) increased the odds of belonging to a genotypic cluster in relation to request status. These results indicate a possible shift in reasoning behind genotype requests in 2013. Clinicians were likely recognizing that genotyping provides a deeper understanding of the molecular epidemiology of TB and were thus issuing genotyping requests not only to address a specific hypothesis about transmission but also to understand the overall transmission dynamics of TB in BC.

Prospective universal genotyping will enable earlier detection of clusters and allow prompt intervention.³¹³ However, this can only occur if those capable of acting on the information have timely access to it. Universal genotyping requires an efficient and effective means of communicating genotyping results, such as the online tools developed in other jurisdictions.^{307,319} While implementation of a universal genotyping program incurs additional costs, it is believed

that the incremental expenditure associated with additional genotyping and the cost of implementing a new reporting system are minimal on the scale of a provincial public health budget. This is especially true when considered in the context of TB infections prevented, as the average cost of treating a person with active TB in Canada is \$47,000,¹¹ and when universal genotyping refutes suspected transmission and large-scale contact tracing and case finding are avoided, especially in complex settings such as homeless shelters.³¹³ Tangible benefits are also realized when specimen cross-contamination events are revealed by universal genotyping and an individual can be taken off unnecessary therapy.^{320,321}

While the data make a strong case for implementing universal genotyping in a low-incidence setting, it is impossible to know with certainty how many new infections would have been prevented if universal genotyping had been in place since 2009; thus, the true public health impact of this intervention cannot be fully assessed. However, universal genotyping of *M. tuberculosis* in New York City revealed new transmission sites and contributed to the rapid diagnosis and treatment of both active cases and infected contacts.³¹³ It is also difficult to assess the future potential of universal genotyping in well-resourced settings, where WGS is supplanting MIRU-VNTR genotyping as the method of choice for inferring transmission. Until WGS of all isolates is routinely performed, MIRU-VNTR genotyping and other molecular methods provide valuable insight into a region's TB epidemiology and permit comparison of patterns across jurisdictional boundaries. If countries like Canada are to achieve the ambitious elimination targets set by the World Health Organization, every available tool in our arsenal must be used to accelerate progress toward making TB an infection of the past.

Chapter 3: Molecular Epidemiology of Tuberculosis in British Columbia, Canada—A 10-Year Retrospective Study

3.1 Background

As tuberculosis (TB) prevention and care programs in low-incidence, well-resourced settings look to accelerate progress towards elimination, it's clear that different interventions are required for different populations—whether it be enhanced screening and uptake of latent tuberculosis infection (LTBI) preventative therapy or interventions aimed at accelerating diagnosis and reducing person-to-person transmission. To identify discrete groups of individuals with TB and ultimately develop tailored interventions bespoke to each, molecular genotyping methods such as 24-locus mycobacterial interspersed repetitive unit–variable-number tandem repeat (MIRU-VNTR) can be leveraged.⁷⁷ MIRU-VNTR is a PCR-based technique with high discriminatory power, often used to differentiate relapse from reinfection, detect laboratory cross-contamination events, and identify outbreaks and endemically circulating strains.³²²

Canada has a low TB incidence rate of 4.4 cases per 100,000 population, but amongst the provinces, British Columbia (BC) has one of the highest rates—6.3 cases per 100,000 population.⁹ More than 80% of BC's TB diagnosed individuals live in the Greater Vancouver Region (GVR),⁹ home to approximately half of BC's residents and the majority of BC's immigrant population.³²³ This latter group represents 81% of BC's TB diagnoses,⁹ in whom active TB disease is generally thought to result from reactivation of LTBI acquired in the individual's country of origin. Risk factors for TB disease in this group are likely markedly different from those in the group whose disease results from a locally transmitted infection.

Previous population-based molecular epidemiological studies in Canada have focused largely on specific metropolitan areas,^{51,324,325} with few province-wide studies,^{326–328} and no provincial study has used 24-locus MIRU-VNTR; thus, a retrospective genotypic survey of all culture-

positive TB diagnoses in BC from 2005–2014 was undertaken to better understand the patterns underlying TB transmission in BC.

3.2 Materials and Methods

3.2.1 Study setting and design

The British Columbia Centre for Disease Control (BCCDC)'s Public Health Laboratory (BCPHL) receives all *Mycobacterium tuberculosis* (*Mtb*) cultures for the province, and oversees routine diagnosis, and phenotypic drug sensitivity testing. Prior to 2014, genotyping was performed on request, with approximately 20% of isolates genotyped annually. Therefore, a retrospective study was designed to include all persons with culture-confirmed TB (79.5% of all 2,915 diagnoses), residing in BC whose first *Mtb* isolate was received by the BCPHL from 2005 through 2014 ($n = 2,318$). *Mycobacterium africanum* ($n = 29$), *Mycobacterium bovis* ($n = 3$), and *Mycobacterium bovis* bacilli Calmette-Guérin ($n = 19$) were excluded from the analysis—these are not commonly isolated at BCPHL and local transmission was not expected. For individuals with a recurrence during the study period, data from their first episode only was used if isolates from their first and second episode had matching MIRU-VNTR ($n = 11$), and data from both episodes where MIRU-VNTR indicated reinfection ($n = 2$).

3.2.2 Case data

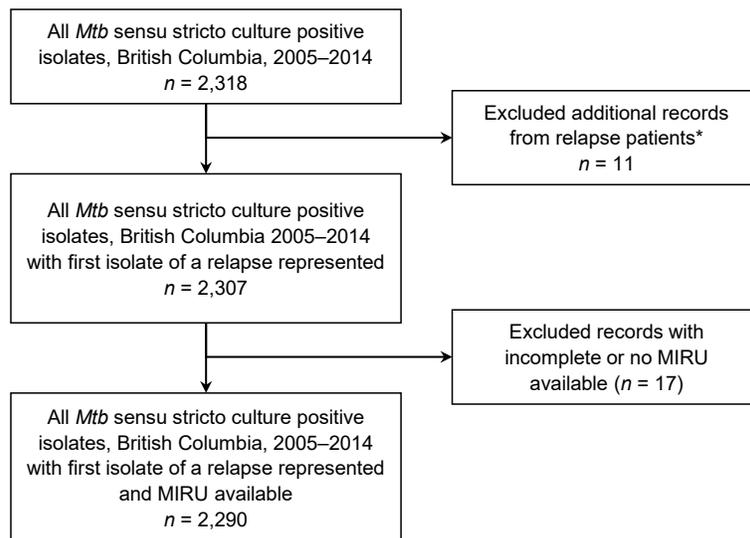
Individual-level clinical and demographic data were extracted from BCCDC's Integrated Public Health Information System (iPHIS), which contains both routinely collected surveillance data as reported to the Public Health Agency of Canada and additional data collected in the course of epidemiological contact investigations. Community type was determined using the population density of the geographic service area in which each individual resided—metro ($>190,000$), urban/rural ($40,001–190,000$), rural ($10,001–40,000$), and remote ($\leq 10,000$). Postal codes were used to obtain the corresponding census dissemination area (DA) for each case and linked it to the 2006 Canadian Marginalization Index (CAN-Marg)³¹⁶ to determine the deprivation index quintile—a neighbourhood-level indicator of socioeconomic status. The CAN-Marg material

deprivation index measures relative socioeconomic disadvantage of a DA compared to the rest of Canada, reported as quintile values by DA (quintile 1: least deprived, quintile 5: most deprived).

3.2.3 Laboratory analysis

All *Mtb* isolates were revived from BCPHL’s frozen archival stocks on Lowenstein-Jensen (LJ) slants or in MGIT™ liquid medium (Becton-Dickinson, Sparks, MD). Phenotypic drug susceptibility results (isoniazid, rifampin, ethambutol, and streptomycin) were available for each isolate from routine testing on the BACTEC™ MGIT™ 460 or 960 (Becton-Dickinson, Sparks, MD). DNA was extracted using the MagMAX™ Total Nucleic Acid Isolation Kit (Ambion, Austin, TX).

Of the 2,307 culture-positive isolates meeting study criteria (**Figure 3-1**), 17 isolates had incomplete MIRU-VNTR or were unavailable for genotyping—leaving a total of 2,290 (99.3%) isolates which were successfully genotyped using standard methods.⁷⁷ Major lineage was predicted for each isolate using TB-Insight’s CBN method.³²⁹ Phylogenetic relationships within each lineage were visualized using a minimum-spanning tree (MST) in PHYLOViZ 2.0.³³⁰



*First episode for a relapse patient was maintained in the study; relapse was defined as a subsequent episode with a genotype ≤ 1 MIRU-VNTR locus different to the initial episode.

Figure 3-1. Molecular epidemiology study inclusion/exclusion criteria.

3.2.4 Statistical analysis

A cluster was defined as ≥ 2 isolates with identical MIRU-VNTR patterns. The odds ratio (OR) and 95% confidence interval (CI) was then estimated for the distribution of persons by cluster status (clustered/non-clustered) according to birthplace and other clinical and demographic variables. To examine factors associated with cluster growth a multivariable logistic regression model was constructed with cluster size—large (≥ 10 persons) versus small (< 10 persons)—as the outcome, using backward elimination of factors identified in univariable analysis ($p < 0.20$), and Akaike's Information Criterion (AIC) minimisation.³³¹ Because the variables (HIV status, illicit drug and alcohol misuse) had $> 5\%$ missing values, Little's test³³² was performed to assess whether these data were missing completely at random (MCAR). The test suggested no violation of this assumption and missing values were unrelated to genotypic clustering ($p > 0.05$). To test the association between TB lineage and disease site, a Chi-square test was used, and to examine time from immigration to active TB disease, as well as median age between clustered and non-clustered individuals, the Mann-Whitney U test was used. All analyses were executed in R (v3.3.1).

3.3 Results

Table 3-1 presents an overview of the demographic and clinical characteristics of culture-positive tuberculosis in BC. The median age was 52 years, with the highest proportion of diagnoses occurring in individuals aged 35–54. Males outnumbered females 1.4:1. Country of birth was available for 97.5% of individuals, most of whom (73.7%) were non-Canadian-born. Although 78 countries were represented, most TB cases born outside Canada came from high-incidence settings,³³³ with 23.2% from India, 20.9% from Philippines, 18.5% from China, and 25.0% from other Asian countries. Most individuals (76.6%) lived in metro regions at the time of TB diagnosis. With respect to clinical characteristics, most (77.2%) had respiratory TB, and of these 16.3% of individuals were characterized as having cavitary disease based on chest radiography. Of the persons for whom HIV status was known (82.4%), 103 individuals were HIV-positive. A small fraction of individuals were recorded as using drugs (5.7%) or alcohol (5.5%).

Phenotypic drug susceptibilities were available for all genotyped isolates, with multi-drug resistance (MDR) defined as resistance to at least isoniazid and rifampin found in 18 (0.8%) isolates (**Table 3-2**).

Table 3-1. Study population. Demographic and clinical characteristics of culture positive TB cases, British Columbia 2005–2014 ($n = 2,290$).^a

Characteristic	No. Cases (%)
Age, years	
0–14	32 (1.4)
15–34	500 (21.8)
35–54	704 (30.7)
55–74	584 (25.5)
75+	470 (20.5)
Gender ^b	
Male	1329 (58.1)
Community type	
Metro	1753 (76.6)
Urban/Rural	332 (14.5)
Rural	173 (7.6)
Remote	32 (1.4)
Birthplace ^c	
Canada	588 (26.3)
Non-Canadian-born continent ^d	
Asia	1437 (87.6)
Africa	79 (4.8)
Europe	69 (4.2)
Americas	45 (2.7)
Oceania	11 (0.7)
Time in Canada ^e	
< 5 years	456 (28.6)
≥ 5 years	1141 (71.4)
Disease Site	
Respiratory	1767 (77.2)
Non-Respiratory	363 (15.9)
Respiratory + Non-Respiratory	160 (7.0)
Respiratory ^f smear	
Positive	1152 (62.1)
Cavitary disease	
Yes	315 (13.8)
Drug susceptibility	
MDR	18 (0.8)
INH-R (non-MDR)	173 (7.6)

Continued on next page

Table 3-1 *Continued from previous page*

Characteristic	No. Cases (%)
HIV	
Positive	103 (4.5)
Negative	1784 (77.9)
Unknown	403 (17.6)
Illicit drug use	
Yes	130 (5.7)
No	1639 (71.6)
Unknown	521 (22.8)
Alcohol misuse	
Yes	125 (5.5)
No	1656 (72.3)
Unknown	509 (22.2)
Material deprivation ^g	
Quintile 1 (least)	273 (12.5)
Quintile 2	418 (19.2)
Quintile 3	529 (24.3)
Quintile 4	529 (24.3)
Quintile 5 (most)	427 (19.6)

Abbreviations: HIV, human immunodeficiency virus; INH-R, isoniazid resistant; MDR, multidrug-resistant (tuberculosis resistant to isoniazid and rifampin).

^aPercentages have been rounded and may not total 100%.

^bOne transgender/gender-unknown individual excluded from analysis.

^cData unavailable for 57 individuals.

^dData unavailable for 4 individuals.

^eData unavailable for 48 individuals.

^f“Other respiratory” sites (e.g. pleura) were excluded.

^gData unavailable for 114 individuals.

Table 3-2. Multi-drug resistant isolates. Characteristics of *Mycobacterium tuberculosis* multi-drug resistant cases in British Columbia, 2005–2014 ($n = 18$).

Birth Sub-Continent	Region	Case Type	Disease Site	Lineage	MIRU-VNTR Cluster ^a
North America	GVR	New	Resp.	East-Asian	Yes
North America	GVR	New	Resp.	East-Asian	Yes ^b
East Asia	GVR	New	Resp.	East-Asian	Yes ^b
East Asia	GVR	New	Resp.	East-Asian	
East Asia	GVR	Retreatment	Resp.	East-Asian	Yes
East Asia	GVR	Retreatment	Resp.	East-Asian	
South-Central Asia	GVR	New	Resp.	East-African-Indian	Yes
South-Eastern Asia	GVR	Retreatment	Resp.	Indo-Oceanic	
South-Eastern Asia	GVR	New	Non-Resp.	Indo-Oceanic	
South-Eastern Asia	GVR	New	Non-Resp.	Indo-Oceanic	Yes
South-Eastern Asia	GVR	New	Resp.+Non-Resp.	Indo-Oceanic	
Northern Europe	GVR	New	Resp.	East-Asian	Yes
East Asia	GVR	Retreatment	Resp.	East-Asian	
East Asia	GVR	New	Resp.	East-Asian	Yes
East Asia	non-GVR	Retreatment	Resp.	East-Asian	
East Asia	GVR	Retreatment	Resp.	East-Asian	
South-Eastern Asia	GVR	New	Resp.	East-Asian	
South-Eastern Asia	GVR	New	Resp.	East-Asian	

Abbreviations: GVR, Greater Vancouver Region; Non-Resp., non-respiratory; Resp., respiratory.

^aYes indicates that the isolate belongs to a MIRU-VNTR cluster.

^bSame cluster; known transmission event.

3.3.1 Lineage analysis

First, the phylogenetic structure of BC's *Mtb* population was examined and the association between lineage and the study variables explored. An MST revealed numerous large Euro-American clusters with distinct clades containing sizable clusters (**Figure 3-2**). Consistent with previous research,⁴¹ it was found that lineage reflected birthplace (**Figure 3-3**), and the Euro-American group contained largely Canadian-born persons (57.7%). The majority (13/18) of MDR isolates belonged to the East-Asian lineage (**Table 3-2**). Disease site varied by lineage, and it was found that the proportion of exclusively non-respiratory tuberculosis was higher amongst individuals with an Indo-Oceanic lineage (26.7%) versus other lineages: East-Asian Indian (18.2%), East-Asian (12.6%), and Euro-American (10.4%), $p < 0.001$. Persons with an Indo-Oceanic strain also had the highest proportion of respiratory disease with non-respiratory involvement (**Table 3-3**).

Clustering rates varied between lineages, with 54.5% of Euro-American, 43.3% of East-Asian, 33.8% of Indo-Oceanic and 22.7% of East-African Indian isolates clustering. The five largest clusters belonged to the Euro-American lineage (**Table 3-4**).

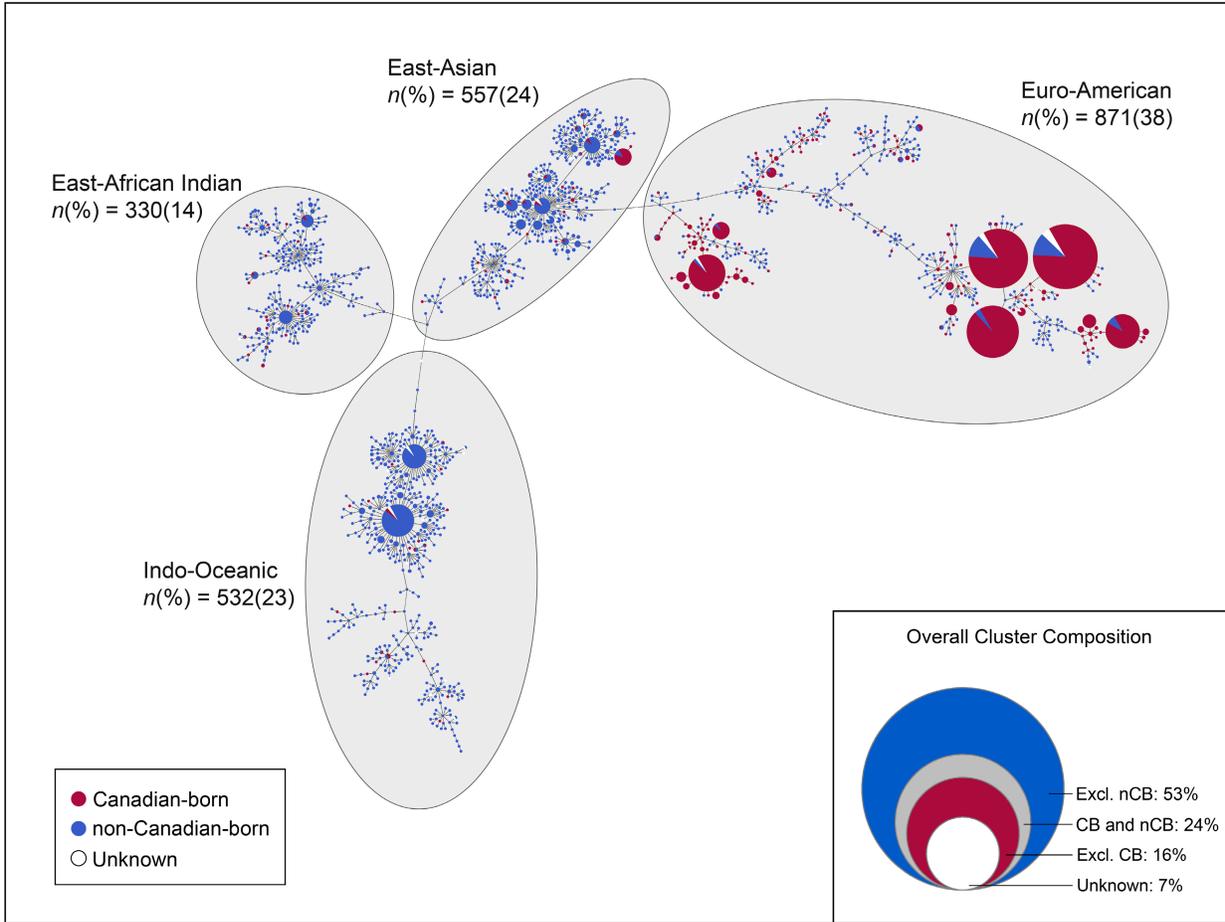


Figure 3-2. Population structure of *Mycobacterium tuberculosis* genotypes in BC. Minimum spanning tree analysis of mycobacterial interspersed repetitive-unit-variable-number tandem repeat (MIRU-VNTR) genotyping for *Mycobacterium tuberculosis* isolates, British Columbia (2005–2014). The size of each circle is proportional to the number of isolates. Classification of strains by birthplace is visualized with color coding. The inset demonstrates overall cluster composition with respect to birthplace; relative frequency of clusters that were exclusively Canadian-born (Excl.CB), exclusively non-Canadian-born (Excl. nCB), Canadian- and non-Canadian-born (CB and nCB), or where there were cases in a cluster with only CB or nCB identified in addition to ≥ 1 case of unknown birthplace. Note that percentages have been rounded and may not total to 100%.

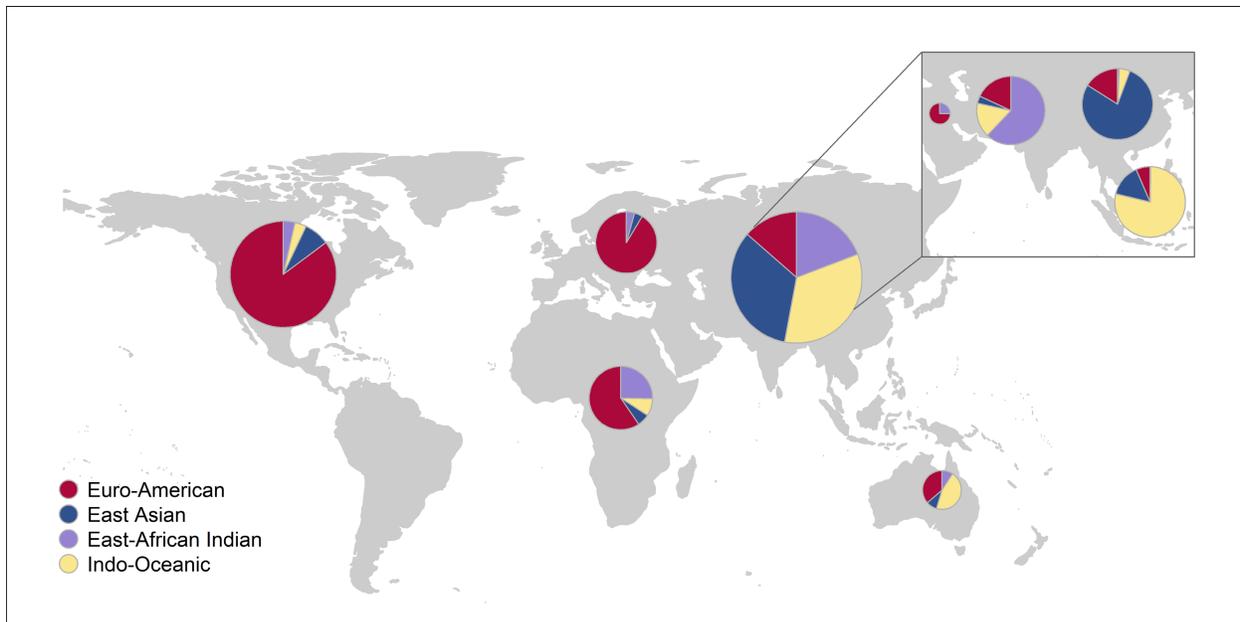


Figure 3-3. *Mycobacterium tuberculosis* lineage by continent of birth. Relative frequency of *Mycobacterium tuberculosis* (*Mtb*) *sensu stricto* lineages by continent of birth of cases. The inset map separates the *Mtb* lineages of Asian-born persons into the relative frequency observed within each Asian sub-continent. Pie chart areas are scaled relative to the number of cases.

Table 3-3. Lineage by anatomical disease site. *Mycobacterium tuberculosis sensu stricto* lineages by anatomical disease site in British Columbia, 2005–2014.^a

Lineage	No. Cases (%)			NRTB vs. RTB OR (95%CI)
	RTB	NRTB	RTB+NRTB	
Euro-American	723 (83.0)	91 (10.4)	57 (6.5)	Reference
East-Asian	456 (81.9)	70 (12.6)	31 (5.6)	1.2 (0.9–1.7)
East-African Indian	252 (76.4)	60 (18.2)	18 (5.5)	1.9 (1.3–2.7)
Indo-Oceanic	336 (63.2)	142 (26.7)	54 (10.2)	3.4 (2.5–4.5)

Abbreviations: CI, confidence interval; NRTB, exclusively non-respiratory; OR, odds ratio; RTB, exclusively respiratory; RTB+NRTB, respiratory+non-respiratory.

^aPercentages have been rounded and may not total 100%.

Table 3-4. Large cluster characteristics. Characteristics of 24-locus MIRU–VNTR clusters comprised of ≥ 10 individuals, displayed as clusters that were predominantly Canadian- or non-Canadian-born: British Columbia, 2005–2014.

Cluster ID	Cluster Size	Predominant Birthplace ^a (%)	Median Age (IQR) years	Gender M:F	Predominant Community Type (%)	Lineage
Canadian-born						
MClust-002	70	Canada (88.1)	50 (43–57)	13.0	Metro (78.6)	Euro-American
MClust-012	64	Canada (87.1)	48 (40–56)	3.0	Metro (79.7)	Euro-American
MClust-001	56	Canada (96.4)	40 (29–48)	0.8	Rural (75.0)	Euro-American
MClust-003	39	Canada (97.4)	45 (29–50)	1.8	Rural (56.4)	Euro-American
MClust-008	36	Canada (91.7)	41 (35–55)	1.1	Metro (83.3)	Euro-American
MClust-035	17	Canada (88.2)	39 (32–58)	1.7 ^b	Metro (82.4)	East-Asian
MClust-052	17	Canada (94.1)	49 (46–53)	3.2	Metro (82.4)	Euro-American
MClust-134	13	Canada (100.0)	59 (46–68)	3.3	Rural (84.6)	Euro-American
MClust-055	10	Canada (100.0)	38 (31–43)	1.5	Urban/Rural (70.0)	Euro-American
Non-Canadian-born						
MClust-011	34	Philippines (97.0)	41 (31–50)	2.1	Metro (76.5)	Indo-Oceanic
MClust-021	25	Philippines (100.0)	50 (31–54)	0.8	Metro (84.0)	Indo-Oceanic
MClust-038	16	China/Hong Kong (87.5)	76 (64–81)	1.3	Metro (100.0)	East-Asian
MClust-187	16	China/Hong Kong (80.0)	66 (47–87)	0.6	Metro (100.0)	East-Asian
MClust-149	13	India (84.6)	59 (34–79)	0.9	Metro (61.5)	East-African Indian
MClust-046	12	India (75.0)	71 (64–78)	2.0	Metro (91.7)	East-African Indian
MClust-032	11	<i>Mixed</i> ^c	54 (48–82)	1.2	Metro (100.0)	East-Asian

Abbreviations: IQR, interquartile range; MIRU–VNTR, mycobacterial interspersed repetitive-unit–variable-number tandem repeat.

^aInformation for birthplace was unknown for 9 individuals; percentage represents those with complete data.

^bNot included in the ratio, 1 transgender/gender unknown individual.

^cPredominantly East Asian and South-East Asian countries.

3.3.2 MIRU-VNTR identifies discrete subgroups amongst BC's TB cases

Next, individual- and community-level risk factors driving clustering in BC were examined. MIRU-VNTR revealed that of 2,290 isolates, 1,319 (57.6%) were unique profiles, likely reflecting LTBI reactivation, while the remaining 42.4% were grouped into 189 clusters (2–70 isolates/cluster), suggesting potential local transmission (**Table 3-5**). Via the “ $n - 1$ ” method⁸⁸ MIRU-VNTR estimated that 782 (34.1%) of infections could have resulted from local transmission. The median age of non-clustered individuals was higher (56 years) compared to clustered (48 years), $p < 0.001$. Amongst males, 44.6% were clustered versus 39.3% amongst females. Other factors for clustering included HIV, illicit drug use and alcohol misuse (**Table 3-6**).

Within the group of Canadian-born persons the majority (77.0%) were in a cluster, while only 30.2% of persons born outside Canada belonged to a cluster; indeed, the odds of belonging to a cluster were 7.8 times higher for Canadian-born persons (95%CI, 6.2–9.6), **Table 3-6**. Interestingly, few (16.4%) clusters were exclusively Canadian-born (**Figure 3-2**). When stratified by birthplace, risk factors for clustering followed similar trends between Canadian- and non-Canadian-born persons; however, the strength of association differed (**Table 3-6**). For example, both Canadian- and non-Canadian-born persons residing in remote communities had increased odds of belonging to a cluster compared to individuals in metro areas, but odds were higher amongst the Canadian-born (3.6. vs 1.7). Illicit drug use and alcohol misuse were also significantly associated with clustering in Canadian-born persons, and those living in areas of high material deprivation had 2.3 higher odds of belonging to a cluster (95%CI, 1.2–4.4).

Table 3-5. Genotype cluster sizes. Genotyping results (24-locus MIRU-VNTR), including genotype clusters^a ($n = 189$) by size and frequency in British Columbia, 2005–2014.^b

Characteristic	Number	Percentage (%)
<i>Isolates</i>		
Unique genotype	1319	57.6
Clustered genotype	971	42.4
<i>Clusters</i>		
Cluster Size		
2 isolates	102	54.0
3 isolates	33	17.5
4 isolates	7	3.7
5–9 isolates	31	16.4
10–29 isolates	10	5.3
30–49 isolates	3	1.6
≥50 isolates	3	1.6

Abbreviation: MIRU-VNTR, mycobacterial interspersed repetitive-unit–variable-number tandem repeat.

^aClusters are defined as ≥ 2 individuals with *Mycobacterium tuberculosis* infection who share an identical genotype.

^bPercentages have been rounded and may not total 100%.

Table 3-6. Risk factors for genotypic clustering. Distribution and univariable analysis of risk factors associated with *Mycobacterium tuberculosis* genotypic clustering stratified by birthplace, British Columbia 2005–2014.^a

Characteristic	Clustered <i>n</i> (%)	Unique <i>n</i> (%)	All Cases	Canadian-born	Non-Canadian-born
			Clustered vs. Unique OR (95%CI)	Clustered vs. Unique OR (95%CI)	Clustered vs. Unique OR (95%CI)
Age, years					
0–14	16 (50.0)	16 (50.0)	1.3 (0.6–2.6)	0.8 (0.3–2.5)	0.7 (0.2–2.3)
15–34	221 (44.2)	279 (55.8)	Reference	Reference	Reference
35–54	370 (52.6)	334 (47.4)	1.4 (1.1–1.8)	2.4 (1.4–4.2)	1.0 (0.7–1.3)
55–74	237 (40.6)	347 (59.4)	0.9 (0.7–1.1)	0.9 (0.5–1.5)	0.9 (0.6–1.1)
75+	127 (27.0)	343 (73.0)	0.5 (0.4–0.6)	0.3 (0.2–0.6)	0.6 (0.5–0.9)
Gender					
Male	593 (44.6)	736 (55.4)	1.2 (1.1–1.5)	1.1 (0.7–1.6)	1.1 (0.9–1.4)
Female	377 (39.3)	583 (60.7)	Reference	Reference	Reference
Community type					
Metro	678 (38.7)	1075 (61.3)	Reference	Reference	Reference
Urban/Rural	142 (42.8)	190 (57.2)	1.2 (0.9–1.5)	0.7 (0.4–1.1)	0.9 (0.7–1.3)
Rural	126 (72.8)	47 (27.2)	4.3 (3.0–6.0)	2.1 (1.2–3.6)	0.8 (0.4–1.8)
Remote	25 (78.1)	7 (21.9)	5.7 (2.4–13.2)	3.6 (0.8–15.5)	1.7 (0.4–7.7)
Birthplace					
Canada	453 (77.0)	135 (23.0)	7.8 (6.2–9.6)	–	–
Outside Canada	497 (30.2)	1148 (69.8)	Reference		
Disease Site					
Resp.	776 (43.9)	991 (56.1)	1.5 (1.2–1.9)	1.7 (0.9–3.3)	1.0 (0.8–1.3)
Non-Resp.	125 (34.4)	238 (65.6)	Reference	Reference	Reference
Resp. + Non-Resp.	70 (43.8)	90 (56.2)	1.5 (1.0–2.2)	2.1 (0.8–5.9)	1.2 (0.7–1.8)
Respiratory^b smear					
Positive	521 (45.2)	631 (54.8)	1.1 (0.9–1.4)	1.6 (1.0–2.4)	0.9 (0.7–1.1)
Cavitary disease	156 (49.5)	159 (50.5)	1.4 (1.1–1.8)	0.8 (0.5–1.4)	1.3 (1.0–1.8)
HIV positive	66 (64.1)	37 (35.9)	2.6 (1.7–3.9)	1.6 (0.8–3.1)	0.6 (0.3–1.5)
Illicit drug use	112 (86.2)	18 (13.8)	10.3 (6.2–17.0)	2.7 (1.5–5.0)	3.8 (0.9–16.1)
Alcohol misuse	97 (77.6)	28 (22.4)	5.6 (3.6–8.6)	2.7 (1.4–5.1)	1.4 (0.6–3.2)
Material deprivation					
Quintile 1 (least)	100 (36.6)	173 (63.4)	Reference	Reference	Reference
Quintile 2	148 (35.4)	270 (64.6)	0.9 (0.7–1.3)	2.0 (0.9–4.4)	0.9 (0.6–1.4)
Quintile 3	196 (37.1)	333 (62.9)	1.0 (0.8–1.4)	1.3 (0.7–2.6)	1.0 (0.7–1.5)
Quintile 4	220 (41.6)	309 (58.4)	1.2 (0.9–1.7)	1.8 (0.9–3.7)	1.2 (0.8–1.8)
Quintile 5 (most)	224 (52.5)	203 (47.5)	1.9 (1.4–2.6)	2.3 (1.2–4.4)	1.1 (0.7–1.7)

Abbreviations: CI, confidence interval, HIV, human immunodeficiency virus; Non-Resp., non-respiratory; OR, odds ratio; Resp., respiratory.

^aPercentages have been rounded and may not total 100%.

^b“Other respiratory” sites (e.g. pleura) were excluded.

3.3.3 MIRU-VNTR identifies drivers of large transmission clusters

Finally, analyses were conducted to explore the differences between large clusters, typically representing outbreaks requiring public health intervention, from smaller clusters. Individuals in large clusters (≥ 10 persons) were more likely to be Canadian-born (adjusted OR [aOR] 3.3, 95%CI: 2.3–4.8), reside in a rural area (aOR 2.3, 95%CI: 1.2–4.5), or use drugs (aOR 2.0, 95%CI: 1.2–3.4), **Table 3-7**.

Amongst the 16 large clusters (**Table 3-4**), nine were comprised predominantly of Canadian-born individuals ($\geq 87.1\%$), and the few non-Canadian-born individuals within these clusters had a median time from immigration to active TB disease of 40 years (IQR: 25–49). Additionally, for these non-Canadian-born persons, where country of birth was known ($n = 24$), only five (20.8%) individuals emigrated from high-burden TB countries.³³³ Conversely, amongst the seven large clusters comprising mainly non-Canadian-born individuals, most (89.9%) emigrated from high-burden countries and had a significantly lower median time from immigration to active disease (12 years, IQR: 3–18), $p < 0.001$.

Table 3-7. Risk factors associated with cluster size. Multivariable analysis of factors associated with large and small 24-locus MIRU-VNTR clusters in British Columbia 2005–2014 ($n = 971$).^a

Characteristic	Large vs. Small OR (95%CI)	Large vs. Small aOR ^b (95%CI)
Age, years		
0–14	0.9 (0.3–2.6)	0.7 (0.2–2.6)
15–34	Reference	Reference
35–54	1.4 (1.0–2.0)	1.2 (0.8–1.8)
55–74	0.9 (0.6–1.3)	1.1 (0.7–1.8)
75+	0.5 (0.3–0.8)	0.9 (0.5–1.6)
Gender		
Male	1.3 (1.0–1.7)	1.4 (1.0–1.9)
Community type		
Metro	Reference	Reference
Urban/Rural	1.4 (0.9–2.0)	0.9 (0.6–1.5)
Rural	3.2 (2.1–4.9)	2.3 (1.2–4.5)
Remote	0.7 (0.3–1.5)	0.5 (0.2–1.4)
Birthplace		
Canada	4.6 (3.5–6.1)	3.3 (2.3–4.8)
Illicit drug use	4.9 (3.1–7.8)	2.0 (1.2–3.4)

Abbreviations: aOR, adjusted odds ratio; CI, confidence interval; MIRU-VNTR, mycobacterial interspersed repetitive-unit–variable-number tandem repeat; OR, odds ratio.

^aLarge clusters were defined as ≥ 10 persons; small clusters, as < 10 persons.

^bAdjusted for age, gender, community type, birthplace, and illicit drug use.

3.4 Discussion

The molecular epidemiology of tuberculosis in British Columbia from 2005 through 2014 was described and demonstrates using a near-complete (99.3%) isolate collection, that BC has notable strain diversity, with $> 1,500$ distinct MIRU-VNTR genotypes. The *Mtb* population structure reflects the global nature of BC’s residents. Migration to BC has been occurring for several centuries, first by predominately European settlers and later with individuals from all over the world—especially from Asia³³⁴—which is reflected in the proportion of each lineage by region of birth. Clustering rates vary between lineages, with the largest clusters belonging to the Euro-American lineage, typical of what has been reported in European and North American-born

populations.⁴¹ An MST revealed sizable clusters within the Euro-American lineage, and distinct sub-groups, likely reflecting a long history of migration to Canada and independent introduction of strains which have diversified and now circulate endemically, such as those introduced during the fur trade in previous centuries.⁹²

Different *Mtb* lineages have frequently been associated with phenotypic differences such as propensity for drug resistance, varying pathogenicity, and tendencies towards specific disease sites.^{46,335} Indeed, it was observed that the bulk of MDR disease occurring in individuals with East-Asian strains, while individuals with Indo-Oceanic and East-African Indian lineages had higher odds of non-respiratory disease and the lowest clustering rates, an observation in line with a large U.S. study.⁴⁶ Given that non-respiratory TB requires a high index of suspicion for diagnosis and commonly results in diagnostic delays and increased morbidity and mortality,⁶ the observation here suggests that clinicians treating individuals who have emigrated from countries where Indo-Oceanic and East-African Indian strains circulate, might benefit from educational initiatives urging them to “think TB”.

Overall, 189 clusters comprising 42.4% of the study isolates were identified. Clustering rates previously reported from smaller studies in BC have varied substantially, with earlier work in the Greater Vancouver area reporting a much smaller clustering rate of 17.3%,³²⁴ and a study of Western Canadian provinces suggesting clustering from 0–82%.³²⁸ Given the near complete sampling over a decade-long period that was undertaken, the figure reported here represents the most accurate estimate of genotype-level clustering for this setting. Using the “ $n - 1$ ” method,⁸⁸ it was estimated that 34.1% of BC’s cases may be the result of local transmission, a figure identical to a study¹²¹ from London, England—a city with a similarly large and ethnically diverse population. This is certainly an overestimation—reports directly comparing MIRU-VNTR to whole genome sequencing (WGS) have shown that genotype-level identity does not always correspond to genomic distances that reflect recent, local transmission.^{140,336} Indeed, two large Indo-Oceanic clusters were noted whose MIRU-VNTR patterns match those of clusters reported elsewhere in Canada.¹⁴⁰ WGS yielded genomic distances incompatible with local transmission, suggesting that these clusters likely represent regionally endemic strains acquired in the country

of origin rather than transmission within Canada.¹⁴⁰ Subsequent work (**Chapter 8**) includes sequencing of clustered isolates identified here to further refine the estimate of transmission, and will allow us to prioritise MIRU-VNTR clusters for investigation.

Where MIRU-VNTR is most likely to capture true local transmission is amongst the Canadian-born. These individuals had nearly eight times the odds of belonging to a cluster and multiple large clusters were identified—two already characterized by WGS,^{74,144,174} and most known to public health personnel and involving documented epidemiological links. In a New York City–area study, U.S.-born residents were more likely to be involved in transmission clusters compared to cases born outside the U.S., with the authors concluding that transmission occurs almost exclusively within the American-born population.³³⁷ However, in this study, nearly one-quarter of clusters involved both Canadian- and non-Canadian-born individuals, suggesting transmission likely occurs both across and within these populations. A 2014 systematic review of European TB found the percentage of cases in “mixed” clusters ranged from 0% to 34.2%,³³⁸ the extent to which this is occurring in BC will be revealed through genomic investigation.

Understanding where and amongst whom transmission is occurring permits targeted contact tracing and cluster investigation efforts, improved resource allocation, and interventions tailored to local epidemiology. Here, it was found that while incidence was higher in metro areas, the odds of clustering were higher and cluster size was larger in rural and remote settings (**Table 3-6, Table 3-7**), suggesting local transmission dominates in low-density settings, while both local transmission and LTBI reactivation contribute to TB case counts in urban areas. Individual-level factors, including HIV, illicit drug use, alcohol misuse, or residing in a marginalized area, were all associated with increased odds of clustering, consistent with other studies.^{117,339,340} All of this information could be used to develop a risk score for an individual contributing to onward transmission, based on both clinical and demographic factors and a strain’s specific genotype and lineage. Such a score could be used to prioritize individuals for enhanced contact tracing and follow-up during therapy. Additionally, the observation that nearly 70% of non-Canadian-born persons have a unique genotype suggests that targeted LTBI screening is an important strategy for preventing the reactivation that is contributing to the bulk of TB diagnoses in BC.

The dataset included only a small number of MDR-TB cases, the majority of which occurred in non-Canadian-born individuals with East-Asian lineage isolates—a lineage known for its association with drug resistance.³⁴¹ With one exception—a known family transmission—MDR-TB isolates did not cluster by MIRU-VNTR, indicating that transmission most likely occurred prior to arrival in Canada. As immigrant numbers continue to rise in BC, many arriving from regions with high MDR-TB rates, BC is at risk of increased MDR-TB as reported in other low-incidence settings.³⁴² Thus, it is vital to have the molecular tools available to monitor the presence of drug-resistant strains and differentiate MDR-TB resulting through treatment failure from newly acquired MDR-TB infection.

The present study does have some important limitations. As noted, although the discriminatory power of MIRU-VNTR is similar to that of restriction fragment length polymorphism (RFLP),³⁴³ it does not provide the necessary resolution to differentiate closely related isolates, particularly for non-Euro-American lineages.^{140,336} It has been suggested that Euro-American strains were over-represented during method development, leading to a bias in the discriminatory power towards this lineage.³¹⁸ WGS can improve this resolution, which was subsequently undertaken and detailed in **Chapter 8**. Second, an individual's country of birth may not accurately reflect their movement. While *Mtb* lineage often matches what one would expect based on birthplace, some individuals may have lived in other countries prior to arrival in Canada, and furthermore may travel after immigration. Additionally, some of the Canadian-born persons may have non-Canadian-born parents, potentially increasing their risk of TB infection through household exposures and/or travel to their parents' birthplace. This may account for some of the mixed Canadian-/non-Canadian-born clusters. Unfortunately, this level of detail is not often included in most public health databases, precluding its analysis, but these scenarios are likely infrequent. What is clear is that persons born outside Canada is too broad a category and a more refined definition would benefit TB surveillance efforts. Long-time residents of Canada with social risk factors comprise a very different group compared to recent immigrants and should be viewed as a distinct group by TB programs. An investigation of this issue is described in **Chapter 9, Section 9.3.1**.

This study provides a benchmark against which BC can measure future progress and offers new insight into the molecular epidemiology of tuberculosis in the province. This knowledge can be used to support new policy and practice as BC moves towards the ultimate goal of TB elimination, whether it be LTBI screening and prophylaxis in the immigrant population or a risk score to stratify individuals' risk of onward transmission. In a setting with declining TB incidence, contact network heterogeneity means that local pockets of transmission will exist,³⁴⁴ and identifying these quickly is critical to elimination efforts. The finding around rural/remote transmission highlights these regions as hotspots for such pockets. It is recommended to conduct better training of rural clinicians around recognizing TB, improving access to screening and treatment services, and the introduction of mobile technologies to facilitate a virtual clinic model.³⁴⁵ Moreover, to limit the spread of infection a lower threshold for extensive contact tracing in these regions is recommended. In conclusion, it is clear that a multi-pronged approach that includes targeted screening, treatment, and contact tracing informed by molecular epidemiology will have the greatest impact on tuberculosis rates in BC.

Chapter 4: Universal Genotyping Reveals Province-Level Differences in the Molecular Epidemiology of Tuberculosis

4.1 Background

Tuberculosis (TB) remains a major public health issue in Canada. Molecular techniques, such as 24-locus mycobacterial interspersed repetitive unit–variable-number tandem repeat (MIRU-VNTR) genotyping have improved understanding of TB epidemiology, and many jurisdictions are adopting routine genotyping of all *Mycobacterium tuberculosis* (*Mtb*) isolates.^{121,346} Within Canada, Ontario—the province with the largest number of TB cases⁶—was an early adopter of universal genotyping, using MIRU-VNTR to genotype the first culture-positive isolate for each case since mid-2007.³⁰⁷ Interpreting these data in the context of linked clinical and demographic information has facilitated both contact tracing and outbreak detection in the province.¹¹⁷ More recently, British Columbia (BC) retrospectively genotyped all first culture-positive isolates since 2005 (see **Chapter 3**), and implemented universal genotyping in 2015.

Together, Ontario and BC represent a substantial burden of disease in Canada, accounting for more than 50% of the nation’s TB cases,⁶ with rates in both settings largely driven by reactivation of latent TB infection (LTBI) in persons born outside the country. Both provinces are popular destinations for immigrants, with the large multi-cultural cities of Toronto and Vancouver attracting many newcomers.³⁴⁷ Vancouver has a high proportion of immigrants from Asia, whereas Toronto is more diverse, and in addition to people from Asia, also has many immigrants from Africa, the Caribbean and Latin America.³⁴⁷ Furthermore, although Ontario and BC are thousands of kilometres apart, there is substantial interprovincial migration, with ~15,000 individuals reportedly migrating between Ontario and BC in 2016/17,³⁴⁸ often for economic reasons or job opportunities. Migrants frequently lack support networks and are at greater risk for homelessness and other factors associated with increased risk of TB reactivation or infection.³⁴⁹ This is particularly true in BC, where under-housed migrants from other provinces—some of whom are experiencing mental illness, addictions, and/or chronic health conditions^{349,350}—are thought to be attracted to Vancouver by the temperate climate.

Each Canadian province/territory works independently towards TB prevention and care and contributes data to the national TB surveillance programs—The Canadian Tuberculosis Reporting System and the Canadian Tuberculosis Laboratory Surveillance System. There is currently no national-level TB molecular surveillance program; however, molecular genotyping data are shared informally between provinces and nationally. With both Ontario and BC now having complete MIRU-VNTR genotyping linked to case-level data dating back over a decade, there is a unique opportunity to compare the molecular epidemiology of TB between the two provinces to provide context to the genotypes observed within each region—thereby improving our understanding of genotypic clustering as it relates to local spread of TB, and investigating the frequency of interprovincial TB transmission.

4.2 Methods

4.2.1 Study setting and design

Ontario and BC are the first and third most populous Canadian provinces, respectively, with 14.2 and 4.8 million inhabitants,³⁴⁸ and rank first and second for the highest population proportion of immigrants, at 28.5% for Ontario and 27.6% for BC.³⁵¹ All *Mycobacterium tuberculosis* (*Mtb*) isolates are either identified in culture at the provincial reference laboratories—Public Health Ontario Laboratory (PHOL) and British Columbia Centre for Disease Control Public Health Laboratory (BCPHL), or submitted for reference testing from other laboratories. The study population included all culture-positive TB cases residing in Ontario or BC at TB treatment initiation, with a first *Mtb sensu stricto* isolate received from 2008 through 2014. Included were 3,314 Ontario and 1,602 BC isolates, representing 75.2% and 79.7% of all notified TB diagnoses during this time period in the respective provinces. For individuals with a reoccurrence during the study period indicative of relapse—successful completion of treatment and identical MIRU-VNTR results for both episodes (Ontario: $n = 5$, BC: $n = 9$), only data from the first episode was included.

4.2.2 Diagnosis and case information

All TB cases diagnosed in Ontario are reported to the responsible public health unit, and in BC are reported to the British Columbia Centre for Disease Control (BCCDC), as well as local public health authorities. Case-level clinical and demographic data such as age, gender, birthplace, and disease site were extracted from the Integrated Public Health Information System (iPHIS) for each province. To assess genotyping in the context of urban/rural regions, community type was determined using Statistics Canada-defined health region Peer Groups (A–I),³⁵² which were grouped into four higher-level categories: Metro (G), Urban, high-density (A, H), Urban, moderate-density (B), Rural/Remote (C–E, I).

4.2.3 Genotyping by 24-locus MIRU-VNTR

Using standard methods,⁷⁷ MIRU-VNTR genotyping was completed for 97.8% (3,314/3,388) of Ontario isolates and 99.8% (1,602/1,605) of BC's. Isolates lacking an amplicon peak at any locus had MIRU-VNTR repeated with newly extracted DNA, and where there remained no peak at a single locus—excluding loci 2163 and 2165, which are known to be absent in some strains³⁵³—the locus was coded as missing data and the isolate included in the analyses. Major lineage was predicted using TB-Insight's conformal Bayesian network (CBN) method.³²⁹ An intraprovincial cluster was defined as ≥ 2 isolates with an identical MIRU-VNTR pattern within a province, and where one or more isolates shared an identical genotype across the two provinces, this was defined this as an interprovincial cluster. Genotypic clusters within each major lineage were visualized using a Minimum Spanning Tree (MST) created in PHYLOViZ 2.0³³⁰ and coloured by province. A circular chord diagram was used to graphically represent the relationship between the number of isolates contributing to a genotype match between the provinces—interprovincial clusters were displayed according to the number of isolates contributing to an interprovincial genotype match: single (1 isolate), small (2–9 isolates), large (≥ 10 isolates).

4.2.4 Statistical analysis

Clinical and demographic characteristics were compared between Ontario and BC using a Chi-square test for categorical variables (Fisher's Exact test where appropriate), and a t-test for continuous variables. Intraprovincial clustering proportions were also compared using Chi-square. Multivariable logistic regression was used to examine factors associated with interprovincial clustering, calculating the odds ratio (OR), adjusted OR (aOR), and 95% confidence interval (CI). To calculate the number of clustered TB cases, i.e. those that were potentially attributable to local transmission, the “ $n - 1$ ” method was used in which the first case of each cluster is assumed to have initiated the cluster and is subtracted from the total number of clustered isolates.⁸⁸ A complete-case analysis strategy (excluded records with missing data: $n = 109$ [2.2%]) was used, with stepwise backward selection of variables following Akaike Information Criterion minimization. All statistical analyses were conducted using R statistical software (v3.4.1).

4.3 Results

4.3.1 Descriptive epidemiology

The study population included a total of 4,916 cases (3,314 in Ontario and 1,602 in BC) with a diagnosis of culture-positive TB from 2008 through 2014. The median age was 46 in Ontario with an interquartile range (IQR) of 30–67—significantly lower than in BC (53 years, IQR: 37–72), $p < 0.001$. Case distribution by community type varied between the provinces (**Table 4-1**), with many Ontario cases residing in Metro areas (47.0%) and most BC cases in high-density urban areas (55.7%). Notably, BC had a higher proportion of cases residing in rural/remote regions (11.8% versus 4.1%). Country of birth was available for 97.5% of individuals, the majority of whom were born outside Canada (**Table 4-1**); however, the proportion varied significantly between Ontario (91.3%) and BC (73.5%). Furthermore, Ontario had a higher proportion of recent immigrants—those arriving within the last five years—($n = 1,024$; 35.5%) compared to BC ($n = 309$; 27.8%). BC had a higher proportion of persons with respiratory disease (85.1%) versus Ontario (74.9%).

Table 4-1. Study population. Demographic and clinical characteristics of culture-positive cases 2008–2014, Ontario ($n = 3,314$) and British Columbia ($n = 1,602$).

Characteristic	Ontario	British Columbia	<i>p</i> -value ^b
	<i>n</i> (%) ^a	<i>n</i> (%) ^a	
Age, years			<0.001
0–14	51 (1.5)	20 (1.2)	
15–34	1001 (30.2)	339 (21.2)	
35–54	952 (28.7)	470 (29.3)	
55–74	747 (22.5)	425 (26.5)	
75+	563 (17.0)	348 (21.7)	
Gender ^c			0.041
Male	1838 (55.5)	939 (58.6)	
Community type			<0.001
Metro	1556 (47.0)	449 (28.0)	
Urban, high-density	1132 (34.2)	893 (55.7)	
Urban, moderate-density	490 (14.8)	71 (4.4)	
Rural/Remote	136 (4.1)	189 (11.8)	
Birthplace ^d			<0.001
Canada	284 (8.7)	412 (26.5)	
Non-Canadian-born continent ^e			<0.001
Asia	2282 (77.2)	1017 (89.0)	
Africa	382 (12.9)	50 (4.4)	
Europe	167 (5.6)	45 (3.9)	
Americas	120 (4.1)	24 (2.1)	
Oceania	5 (0.2)	7 (0.6)	
Time in Canada ^f			<0.001
< 5 years	1024 (35.5)	309 (27.8)	
≥ 5 years	1860 (64.5)	801 (72.2)	
Disease Site ^g			<0.001
Respiratory	2256 (68.1)	1250 (78.0)	
Non-Respiratory	832 (25.1)	238 (14.9)	
Respiratory + Non-Respiratory	226 (6.8)	114 (7.1)	

^aPercentages have been rounded and may not total 100%.

^bChi-square test.

^cData unavailable for 4 Ontario individuals.

^dData unavailable for 60 Ontario and 45 British Columbia individuals.

^eData unavailable for 14 Ontario and 2 British Columbia individuals.

^fData unavailable for 86 Ontario and 35 British Columbia individuals.

^g“Other respiratory” sites (e.g. pleura) were excluded.

4.3.2 TB isolates in BC are more likely to be clustered by MIRU-VNTR

MIRU-VNTR genotyping grouped the Ontario *Mtb* isolates into 290 clusters, with a mean cluster size of four isolates (size range: 2–49), yielding a clustered proportion of 31.8% (**Table 4-2**). In BC, 134 clusters were identified, with an average cluster size of five isolates (size range: 2–68) and an overall clustered proportion of 40.5%—significantly higher than in Ontario ($p < 0.001$). Using the “ $n - 1$ ” method,⁸⁸ the number of infections potentially attributable to local transmission was 1,053 (23.0%) in Ontario and 649 (32.1%) in BC. In both provinces, more than half the clusters—56.7% in Ontario and 54.9% in BC—contained only two individuals, with few large clusters of ≥ 10 individuals in either province (Ontario: $n = 11$ [3.8%], BC: $n = 10$ [7.5%]). Differences in the clustered proportion between the two provinces was largely driven by clustering amongst Canadian-born persons (Ontario: $n = 142$ [50.0%], BC: $n = 312$ [75.7%]), as the clustered proportion was similar for persons born outside Canada (Ontario: $n = 892$ [30.0%], BC: $n = 322$ [28.1%]).

Table 4-2. Genotype cluster sizes. Genotype (24-locus MIRU-VNTR) results, including intraprovincial genotype clustering^a by size and frequency in Ontario and British Columbia, 2008–2014.

Characteristic	Ontario <i>n</i> (%) ^b	British Columbia <i>n</i> (%) ^b
<i>Isolates</i>		
Unique genotype	2261 (68.2)	953 (59.5)
Clustered genotype	1053 (31.8)	649 (40.5)
<i>Clusters</i>		
Cluster Size		
2 isolates	164 (56.7)	73 (54.9)
3 isolates	56 (19.4)	23 (17.3)
4 isolates	20 (6.9)	5 (3.8)
5–9 isolates	38 (13.1)	22 (16.5)
10–29 isolates	9 (3.1)	7 (5.3)
30–49 isolates	2 (0.7)	2 (1.5)
≥50 isolates	0 (0.0)	1 (0.8)

Abbreviation: MIRU-VNTR, mycobacterial interspersed repetitive-unit–variable-number tandem repeat.

^aClusters are defined as ≥2 individuals with *Mycobacterium tuberculosis* infection who share an identical genotype.

^bPercentages have been rounded and may not total 100%.

4.3.3 Interprovincial clustering occurs frequently between Ontario and BC

In total, 3,461 distinct MIRU-VNTR patterns were observed across both provinces. Although only 175 of these patterns were detected in both Ontario and BC (**Figure 4-1**), 22.4% (1,102/4,916) of all study isolates had a genotype pattern detected in both provinces—595 (18.0%) Ontario isolates and 507 (31.6%) BC isolates. The majority of these interprovincially matched isolates were also clustered within their respective provinces—85.5% (509/595) in Ontario and 79.1% (401/507) in BC (**Figure 4-2**).



Figure 4-1. Genotypes shared between provinces. Venn diagram representing the number of unique and shared 24-locus MIRU-VNTR genotypes between Ontario and British Columbia, 2008–2014.

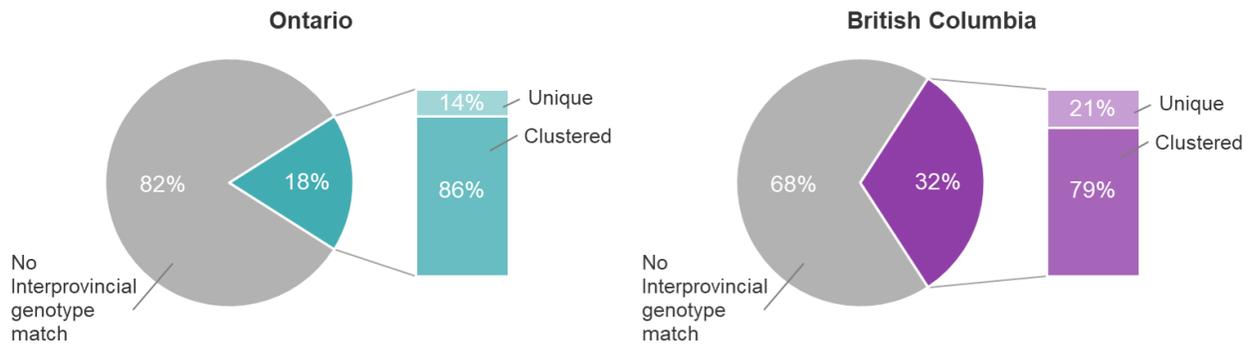


Figure 4-2. Proportion of genotypic clustering. Intra- and interprovincial 24-locus MIRU-VNTR genotypic clustering, Ontario and British Columbia (2008–2014). Each coloured pie wedge represents the proportion of isolates within the province that have a genotype match in the other. For the group that does have an interprovincial match, the stacked bar graphs show the relative frequency of isolates that are clustered or unique within the respective province.

Multivariable logistic regression was used to investigate independent factors associated with interprovincial genotype matches (**Table 4-3**) and found increased odds of matching for BC isolates (aOR 2.1, 95%CI: 1.8–2.5), Canadian-born persons (aOR 2.5, 95%CI: 1.9–3.2), and those with a non-Euro-American lineage *Mtb* isolate (aOR range: 1.9–4.7). Individuals residing in a Metro area had 1.8 times the odds of their isolate belonging to an interprovincial cluster (95%CI: 1.2–2.5) compared to those residing in a rural/remote region. Restricting the sample to include only isolates contributing to an interprovincial genotypic cluster and comparing single versus multiple contributors to a cluster, very similar trends to the factors associated with overall interprovincial clustering were observed (**Table 4-4**). Furthermore, upon examination of cluster composition (**Figure 4-3**), it was found that 68 of the 175 interprovincial clusters were comprised solely of a single isolate detected in each province—80.9% were East-Asian, East-Asian African, or Indo-Oceanic clusters and 93.9% of these isolates were identified in persons born outside Canada, **Figure 4-4**.

Table 4-3. Multivariable logistic regression. Distribution, frequency, and logistic regression analysis of factors associated with interprovincial genotypic clustering of *Mycobacterium tuberculosis* isolates between Ontario and British Columbia 2008–2014 ($n = 4,807$).

Characteristic	Interprovincial Genotype Match		Interprovincial Genotype Match	
	Yes <i>n</i> (%)	No <i>n</i> (%)	Yes vs. No OR (95%CI)	Yes vs. No aOR^a (95%CI)
Total	1075 (22.4)	3732 (77.6)		
Age, years				
0–14	13 (18.3)	58 (81.7)	0.8 (0.4–1.4)	0.6 (0.3–1.1)
15–34	303 (22.8)	1026 (77.2)	Reference	Reference
35–54	357 (25.4)	1049 (74.6)	1.2 (1.0–1.4)	1.1 (0.9–1.3)
55–74	251 (22.0)	889 (78.0)	1.0 (0.8–1.2)	0.9 (0.7–1.1)
75+	151 (17.5)	710 (82.5)	0.7 (0.6–0.9)	0.6 (0.5–0.8)
Gender				
Male	610 (22.6)	2094 (77.4)	1.0 (0.9–1.2)	1.0 (0.9–1.2)
Female	465 (22.1)	1638 (77.9)	Reference	Reference
Province				
Ontario	578 (17.8)	2672 (82.2)	Reference	Reference
British Columbia	497 (31.9)	1060 (68.1)	2.2 (1.9–2.5)	2.1 (1.8–2.5)
Community type				
Metro	458 (23.2)	1518 (76.8)	1.3 (1.0–1.8)	1.8 (1.2–2.5)
Urban, high-density	470 (23.6)	1520 (76.4)	1.3 (1.0–1.8)	1.6 (1.1–2.2)
Urban, moderate-density	92 (16.8)	455 (83.2)	0.9 (0.6–1.3)	1.4 (1.0–2.2)
Rural/Remote	55 (18.7)	239 (81.3)	Reference	Reference
Birthplace				
Canada	192 (27.6)	503 (72.4)	1.4 (1.2–1.7)	2.5 (1.9–3.2)
Outside Canada	883 (21.5)	3229 (78.5)	Reference	Reference
Lineage				
Euro-American	233 (14.4)	1387 (85.6)	Reference	Reference
East-Asian	340 (35.1)	628 (64.9)	3.2 (2.7–3.9)	4.7 (3.7–5.8)
East-African Indian	156 (17.3)	744 (82.7)	1.2 (1.0–1.6)	1.9 (1.5–2.4)
Indo-Oceanic	346 (26.2)	973 (73.8)	2.1 (1.8–2.5)	3.1 (2.5–3.9)

Abbreviations: aOR, adjusted odds ratio; CI, confidence interval; OR, odds ratio.

^aAdjusted for age, gender, province, community type, birthplace, lineage.

Table 4-4. Multivariable logistic regression according to size.
Multivariable analysis of factors associated with single and multi (≥ 2 isolates) contributors to an interprovincial 24-MIRU-VNTR cluster, Ontario and British Columbia 2008–2014.

Characteristic	Multi vs. Single OR (95%CI)	Multi vs. Single aOR^a (95%CI)
Age, years		
0–14	1.4 (0.3–6.3)	1.0 (0.2–4.8)
15–34	Reference	Reference
35–54	1.5 (1.0–2.3)	1.4 (0.9–2.2)
55–74	1.2 (0.8–1.8)	1.2 (0.8–1.9)
75+	1.0 (0.6–1.7)	1.3 (0.7–2.1)
Gender		
Male	1.3 (0.9–1.7)	1.2 (0.8–1.6)
Female	Reference	Reference
Province		
Ontario	Reference	Reference
British Columbia	0.6 (0.5–0.9)	0.6 (0.4–0.8)
Community type		
Metro	1.9 (1.0–3.8)	2.7 (1.2–5.8)
Urban, high-density	1.2 (0.6–2.4)	1.9 (0.9–4.1)
Urban, moderate-density	1.7 (0.7–4.0)	2.2 (0.9–5.7)
Rural/Remote	Reference	Reference
Birthplace		
Canada	3.3 (1.9–6.0)	8 (3.8–16.6)
Outside Canada	Reference	Reference
Lineage		
Euro-American	Reference	Reference
East-Asian	0.9 (0.6–1.4)	2.0 (1.2–3.5)
East-African Indian	0.6 (0.4–1.0)	1.4 (0.8–2.6)
Indo-Oceanic	1.0 (0.6–1.6)	2.4 (1.4–4.2)

Abbreviations: CI, confidence interval; OR, odds ratio.

^aAdjusted for age, gender, province, community type, birthplace, lineage.

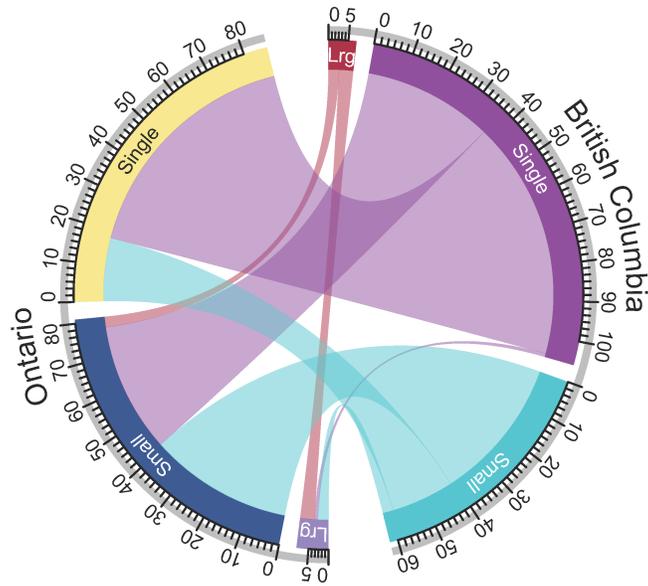


Figure 4-3. Interprovincial genotype matches. A circular chord diagram visualizing the number (indicated by tick marks) of interprovincial 24-locus MIRU-VNTR genotype matches between Ontario and British Columbia from 2008–2014, grouped by the number of isolates within each province sharing the matched genotype: single (1 isolate), small (2–9 isolates), large (≥ 10 isolates). Flow width indicates the number of genotypes.

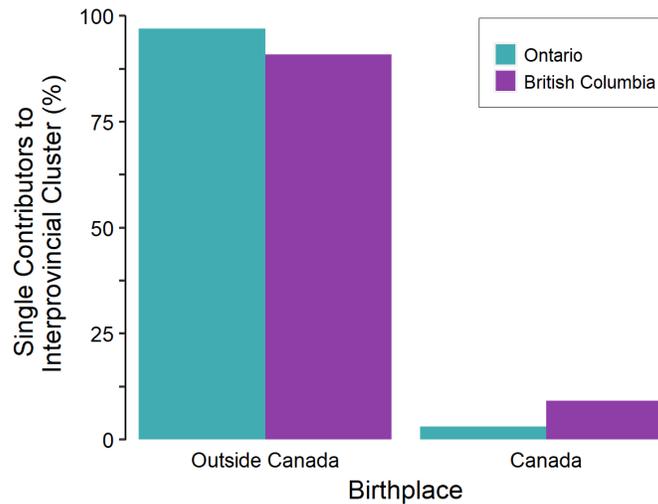


Figure 4-4. Single contributors to clusters. Proportion of single contributors to an interprovincial cluster by province and birthplace.

The 1,894 isolates that were intra- and/or interprovincially clustered were visualized using a minimum spanning tree (**Figure 4-5**), revealing 17 large clusters (≥ 10 persons) across the lineages, many of which were observed in both Ontario and BC. Recognizing that MIRU-VNTR overestimates recent transmission in non-Euro-American lineages,³⁵⁴ and that recent transmission is more likely to occur among Canadian-born individuals, these clusters were examined in the context of lineage and birthplace (**Table 4-5**). Clusters of non-Euro-American lineage isolates were observed in persons born outside Canada, and all but one of these clusters spanned both provinces, suggesting that rather than recent transmission, these clusters may reflect reactivation of strains acquired overseas. Clusters involving predominantly Canadian-born persons tended to occur exclusively within one province or the other and in different community types—Metro and high-density urban in Ontario, largely rural/remote in BC. However, seven isolates with genotypes matching two large BC outbreaks (BC002 and BC012)—one of which has been previously described¹⁷³—appeared in Ontario.

Table 4-5. Large genotypic clusters. Characteristics of 24-locus MIRU–VNTR large clusters (≥ 10 individuals) by predominant birthplace, community type, and lineage: Ontario and British Columbia, 2008–2014.

Cluster ID	Ontario			British Columbia			
	Cluster Size	Predominant Birthplace ^{ab} (%)	Predominant Community Type (%)	Cluster Size	Predominant Birthplace ^{ab} (%)	Predominant Community Type (%)	Lineage
Interprovincial Clusters							
Canadian-born							
ON059/BC002	4	Canada (100.0)	urbanMD/Rural/Remote (75.0)	68	Canada (87.7)	urbanHD (67.6)	EAm
ON065/BC012	3	Canada (50.0)	urbanMD/Rural/Remote (66.7)	46	Canada (91.3)	urbanHD (52.2)	EAm
Non-Canadian-born							
ON253/BC011	49	Philippines (93.8)	Metro (59.2)	30	Philippines (96.6)	urbanHD (56.7)	IO
ON267/BC021	41	Philippines (100.0)	Metro (56.1)	20	Philippines (100.0)	Metro/urbanHD (85.0)	IO
ON155/BC187	26	China (54.2)	Metro (57.7)	15	China (78.6)	urbanHD (60.0)	EAs
ON150/BC038	18	China (61.1)	Metro (55.6)	10	China (100.0)	Metro (60.0)	EAs
ON181/BC046	20	India (65.0)	urbanHD (55.0)	9	India (77.8)	urbanHD (88.9)	EAI
ON012/BC141	19	E. Africa (73.7)	Metro (68.4)	5	E. Africa (80.0)	urbanHD (60.0)	EAI
ON058/–	13	S. Asia (69.2)	Metro (76.9)	1	S. Asia (100.0)	Rural-Remote (100.0)	EAs
ON104/BC157	12	E. Africa/E. Asia (75.0)	Metro/urbanHD (83.3)	4	E. Africa/E. Asia (75.0)	Metro/urbanHD (100.0)	EAs
ON179/BC149	9	India (66.7)	Metro/urbanHD (77.7)	10	India (90.0)	urbanHD (70.0)	EAI
Intraprovincial Clusters							
Canadian-born							
ON219	15	Canada (73.3)	Metro/urbanHD (86.7)	0	–	–	EAm
ON22	14	Canada (76.9)	Metro (64.3)	0	–	–	EAm
BC001	0	–	–	28	Canada (96.4)	Rural/Remote (85.7)	EAm
BC003	0	–	–	26	Canada (96.0)	Rural/Remote (84.6)	EAm
BC008	0	–	–	21	Canada (85.7)	urbanHD (81.0)	EAm
Non-Canadian-born							
ON73	11	E. Africa (55.6)	Metro/urbanHD (72.8)	0	–	–	EAI

Abbreviations: EAm, Euro-American; EAs, East-Asian; EAI, East-Asian Indian; IO, Indo-Oceanic; urbanMD, Urban, moderate-density; urbanHD, Urban, moderate-density.

^aBirthplace was unknown for 10 individuals; percentage represents those with complete data.

^bPredominant birthplace country or region.



Figure 4-5. Population structure of *Mycobacterium tuberculosis* isolates shared between BC and Ontario. Minimum spanning tree analysis of 24-locus MIRU-VNTR of the 1,894 intra- and interprovincially clustered isolates with lineage indicated, Ontario and British Columbia (2008–2014). The size of each circle is proportional to the number of isolates. Classification of genotypes by province is visualized by colour coding.

4.4 Discussion

This study represents the first comprehensive interprovincial comparison of MIRU-VNTR genotyping in Canada using >4,900 *Mtb* isolates collected in Ontario and BC over a seven-year period. This represents >50% of culture-positive TB cases diagnosed in Canada during this period, and provides new insights into the comparative epidemiology of TB in two of Canada's largest provinces, as well as insight into possible interprovincial TB transmission. Although both provinces have large, diverse populations with many people born outside Canada, there were significant differences in the epidemiology and the bacterial population structure between the two provinces. Ontario had more unique MIRU-VNTR patterns, primarily identified within persons born outside Canada, and more cases occurring in large urban areas.

Despite the high strain diversity, the clustered proportion differed significantly between Ontario and BC—similar to findings in a Western Canada study using restriction fragment length polymorphism (RFLP) genotyping in which clustering varied from 9% to 64% across the provinces studied.³²⁸ BC cases were more frequently clustered than those in Ontario, consistent with BC's higher proportion of TB in Canadian-born persons, amongst whom local transmission is likely to drive TB rates. MIRU-VNTR does overestimate true recent transmission,³⁵⁴ so whether these differences in the clustered proportion are still present when whole genome sequencing rather than MIRU-VNTR is used remains to be seen. Encouragingly, most clusters identified with MIRU-VNTR were small, with only seven large outbreaks consistent with recent transmission—most of which have been previously described.^{74,117,173,305} Thus, despite different models of TB management and care between the provinces—with Ontario following a decentralized model and BC a largely centralized system—common practices and national guidelines such as the Canadian Tuberculosis Standards⁶ result in consistently effective public health responses in most cases.

When genotypes present in both provinces were examined, it was found that most MIRU-VNTR matches were due to a single individual in either province, of whom the vast majority were born outside Canada. This is consistent with the notion that these represent LTBI reactivation,³⁵⁵ and although the possibility that these individuals had travelled between Ontario and BC cannot be

excluded, such a transmission scenario is likely rare. There appears to be little interprovincial transmission between Ontario and BC, and the seven cases detected are genotypic matches to two strains endemic to BC circulating within vulnerable populations with known risk factors, including under-housing.^{74,173} It is possible one or more of these Ontario residents had a travel history to or prior residence in BC, with social/behavioural risk factors linked to a higher risk of exposure and infection—something that has been observed in other cross-jurisdictional studies.^{356–358} Interestingly, Ontario’s large MIRU-VNTR clusters (ON219, ON22) circulating amongst under-housed individuals in a metropolitan area of Ontario^{117,305} were not found in BC, suggesting potential differences in the epidemiology or movements of the under-housed populations between the provinces. However, because TB case management in Canada occurs at the provincial/territorial level, sharing of case-level data across jurisdictions is challenging, and prevented the comparison of risk factor and epidemiological data that may have allowed for exploration of within-Canada transmission further. It was also assumed that identical MIRU-VNTR patterns amongst Canadian-born individuals with Euro-American lineage TB represented recent transmission. This observation is supported by recent work in the English Midlands,³⁵⁴ but whether this is the case across thousands of kilometres remains to be seen; it is possible these interprovincial clusters may represent a common strain circulating in Canada amongst vulnerable populations.

Currently, there is no coordinated national molecular surveillance program for tuberculosis in Canada. Although the National Microbiology Laboratory (NML) does offer genotyping services, not all provinces/territories use the service, and instead perform genotyping at their provincial reference laboratory. MIRU-VNTR data, whether generated at NML or at the provincial level, are not routinely shared nationwide, precluding a nationwide molecular surveillance program of the type implemented in the United Kingdom, the Netherlands, and other comparable low-incidence settings.^{75,314} While our analyses suggest minimal TB transmission between BC and Ontario, these are two geographically distant provinces—a similar study using geographically closer jurisdictions may tell a different story.

A national molecular surveillance program is a complex undertaking, requiring coordinated and collaborative efforts by all provinces/territories for implementation, maintenance, support, and evaluation. Perhaps the largest challenge is acquiring funding to support a national program, particularly the necessary personnel required to carry out such an effort, as provincial public health budgets are already limited. Additional issues complicate the ability to access and analyze health data across provincial/territorial borders—data ownership, legal, ethical, and privacy concerns limit what jurisdictions may be willing or able to share, yet these clinical and epidemiological data are required for meaningful interpretation of the genotypic data.³⁵⁹ Interpretation of these data requires trained molecular epidemiologists with a regional- and national-level understanding of TB epidemiology, as well as a suitable information technology platform to link genotyping and administrative data. Laboratory information management systems (LIMS) facilitate the recording and sharing of information at a specimen-level; however, LIMS are not designed for research or surveillance efforts and genotyping data are rarely integrated with the provincial health systems capturing the clinical and epidemiological information for each case. Full integration of these data sources, even within provinces, requires significant resources for creating and curating databases, and routinely linking data. In Ontario, the OUT-TB Web online platform is used to communicate case-level genotyping data across the province and could provide a template for a national system.³⁰⁷ This platform could be expanded to communicate genotyping information at a national level to the appropriate TB program personnel, but careful planning and investment in public health IT infrastructure are needed to ensure a national surveillance program operates as intended and that both personnel and financial resources are available to take action when cross-jurisdictional transmission is detected.

Despite minimal evidence of cross-jurisdictional transmission in the present study, it is believed that the comparison of TB molecular epidemiology between Ontario and BC did further our understanding of local transmission by providing more context to what has been observed in each province independently. Next steps could include expanding the analyses to other Canadian jurisdictions with complete or near-complete genotyping data, as well as incorporating whole genome sequencing data currently being generated in several Canadian settings.

Chapter 5: Genotyping and Whole-Genome Sequencing to Identify Tuberculosis Transmission to Pediatric Patients in British Columbia

5.1 Background

In 2016, there were an estimated 1 million new cases of childhood tuberculosis (TB), causing 253,000 deaths globally.⁵ In low-incidence countries such as Canada, children <18 years have the lowest TB rates of any age group;⁸⁴ however, pediatric cases are often difficult to diagnose and complex to manage. Presentations may be atypical, diagnostic yields are low, and pediatric cases are at greater risk of developing severe, disseminated, and potentially fatal disease without prompt treatment.^{360,361} Tuberculosis in young children is considered indicative of recent transmission. In high-resource regions, reverse contact tracing is usually performed in an attempt to identify and treat the source case to prevent further spread.³⁶²

Contact investigations initiated around pediatric TB cases are most likely to identify a source case in children <2 years, typically revealing an adult caregiver or household member.³⁶²⁻³⁶⁴ However, the epidemiology is not always clear. In a setting such as Canada, immigrant children are less likely to be epidemiologically linked to a known case, and they are often presumed to have acquired infection before immigration.^{87,362,365} Even when a putative source is identified, the molecular epidemiology may not always support a relationship between the *Mycobacterium tuberculosis* (*Mtb*) isolates of the child and presumed source. Genotyping has revealed instances in which a pediatric patient's *Mtb* isolate has a different genotype from their assumed source case, thus refuting transmission,^{362,366,367} and discordant drug-susceptibility patterns have also been used to disprove transmission. Multidrug-resistant (MDR)-TB, particularly in children of immigrants, is believed to result from exposure to adult family members with MDR-TB,^{368,369} yet at least two studies have shown conflicting resistance phenotypes between pediatric cases and their presumed source.^{367,370}

In contrast, concordant genotypes and susceptibility patterns do not necessarily mean that a specific individual was the source of a child's TB infection. Genotyping methods such as mycobacterial interspersed repetitive unit–variable-number tandem repeat (MIRU-VNTR) overestimate clustering in certain lineages of TB that are common among individuals born in high-incidence countries, thus identical genotypes in low-incidence countries often suggest infection with a strain common to a particular ethnic community^{140,336} rather than recent person-to-person transmission—a finding that might influence public health follow-up. The linkages between cases can be further refined through whole-genome sequencing (WGS), which—when combined with epidemiological data—can better identify transmission chains.⁷⁴ Using WGS, *Mtb* isolates with identical genotypes may be separated by enough genomic distance (>5 mutations) to rule out recent transmission from a putative source.¹⁴⁹

To better understand the transmission dynamics of pediatric TB in a low-incidence setting, a retrospective analysis was carried out of all culture-positive TB cases in children <18 years diagnosed in British Columbia (BC), Canada from 2005 through 2014. Routine programmatic contact investigation data combined with MIRU-VNTR and WGS was used to identify source cases and to quantify the extent to which transmission within the province contributes to the overall burden of pediatric TB in BC.

5.2 Methods

5.2.1 Study setting and design

The British Columbia Centre for Disease Control (BCCDC)'s Public Health Laboratory (BCPHL) receives all *Mtb* cultures for the province and performs routine phenotypic drug-susceptibility testing and MIRU-VNTR genotyping on all *Mtb* isolates. Patient care, surveillance, and TB prevention programs are led by Provincial TB Services at the BCCDC. All children <18 years diagnosed with active TB from 2005 through 2014 ($n = 98$) in BC were identified from the provincial surveillance registry and the study population was restricted to only those with a culture-positive diagnosis made in BC ($n = 49$).

5.2.2 Case data

Individual-level clinical, demographic, and contact investigation data were obtained for all study participants through BCCDC's Integrated Public Health Information System (iPHIS). Disease site was categorized as respiratory or non-respiratory.⁶ Information regarding each child's parents' country of origin was obtained through a combination of physician narrative and contact investigation records in iPHIS. Children were categorized by birthplace as non-Canadian-born (nCB) or Canadian-born (CB); the latter group was further subdivided into children born to non-Canadian-born parents (nCBP) or to Canadian-born parents (CBP). Ethnic community was defined by a nCB individual's country of birth, or, for Canadian-born children, their parents' country of birth. In the one case in which the parents were from different countries, the ethnic community of the parent born in a high TB incidence country (≥ 30 cases/100 000)⁶ was used.

5.2.3 Laboratory methods

Mycobacterium tuberculosis isolates were obtained from specimens submitted to the BCPHL for routine testing. Phenotypic drug susceptibility testing results were available all isolates for first-line antibiotics—isoniazid (INH), rifampin (RIF), ethambutol, and streptomycin—with additional data for pyrazinamide in isolates resistant to INH and/or RIF. Isolates were revived from archived stocks, DNA extracted, and genotyped using 24-locus MIRU-VNTR genotyping as described in **Chapter 3**. All 49 (100%) culture-positive isolates were successfully genotyped and isolates whose MIRU-VNTR genotype matched one or more isolates in BC during the study period (see **Chapter 3**) were assigned a cluster identifier (MClustID). All clustered isolates—those from the pediatric cases ($n = 24$) and all isolates from adult cases in each cluster ($n = 202$)—were sequenced using 125 base pairs, paired-end reads on the Illumina HiSeq 2500 platform at the Michael Smith Genome Sciences Centre (Vancouver, BC), according to the following protocol.

TB genomic DNA samples were quantified using Quant-iT assay, and rearrayed in 96-well plates according to input amounts, normalized to 100ng, 30ng, and 1-29ng in 62.5 μ L respectively using JANUS automated workstation (PerkinElmer). The TB genomic DNA was fragmented by Covaris LE220 sonication in a 96 microTUBE Plate for 120 seconds using a “Duty factor” of

30% and “Intensity” of 200 cycles per burst at 450 Peak Incident Power. The paired-end sequencing library was prepared following the BC Cancer Agency’s Genome Sciences Centre’s PCR Enriched 96-well Low Input Small Gap gDNA Library Construction protocol on a Biomek FX robot (Beckman-Coulter, USA) using a customized NEB premix chemistry. Briefly, the DNA was size selected for a 300 bp peak using PCRClean DX beads (0.67:1 to 1:1 beads to sample ratio for upper and lower cut, respectively), and was subject to end-repair, and phosphorylation by T4 DNA polymerase and T4 polynucleotide kinase in a single reaction, followed by purification using PCRClean DX beads (1:1 bead to sample ratio) and 3’ A-tailing by Klenow fragment (3’ to 5’ exo minus). Adenylated libraries were ligated to paired-end adapters using NEB quick ligation premix with enhancer and then subsequently purified twice using PCRClean DX beads (1:1 bead: sample ratio). Adapter ligated templates were PCR-amplified with Phusion DNA Polymerase (Thermo Fisher Scientific Inc. USA) using Illumina’s PE indexed primer set, with the following cycle conditions: 98°C for 60 seconds followed by 6 to 12 cycles of 98°C for 15 seconds, 65°C for 30 seconds and 72°C for 30 seconds, and a final extension at 72°C for 5 minutes. The number of PCR cycles was dependent on the input amount. The PCR products were purified twice using PCRClean DX beads (1:1 bead to sample ratio), and their average size and distributions were determined using Caliper LabChip GX High Sensitivity DNA Chip Assay (PerkinElmer, Inc. USA). Libraries were quantified using Quant-iT High Sensitivity Assay Kit (Invitrogen, CA). Libraries were pooled in equal molar ratio for sequencing on the Illumina HiSeq2500 platform using version 4 chemistry and 125 bp paired end with index reads.

5.2.4 Whole genome sequencing analysis

The resulting FASTQ files were analyzed using a pipeline developed by Oxford University and Public Health England.²⁶⁷ Reads were aligned to the *Mtb* H37Rv reference genome (GenBank ID: NC000962.2), and after masking for low complexity regions an average of 92% of the reference genome was covered. Single nucleotide variants (SNVs) were identified across all mapped nonrepetitive sites. Concatenated SNVs were used to construct maximum-likelihood phylogenetic trees (RAXML 8.2.10,³⁷¹ GTRGAMMA model and 200 bootstrap replicates), which were then viewed using the R statistical software (version 3.4.1). Lineage-defining

SNVs³⁷² were used to classify each sequenced isolate into 1 of the 7 genetic *Mtb* lineages. FASTQ files for all genomes are available at the National Center for Biotechnology Information under BioProject PRJNA413593 and PRJNA49659.

5.2.5 Transmission classification

WGS data were combined with individual-level clinical and epidemiological data, including symptom onset and diagnosis dates, disease site(s), and contact investigation information, to identify the most probable source case for each child's infection. Locally acquired infections were defined as follows: (1) those pediatric cases whose *Mtb* isolate fell within 0–5 SNVs of another isolate from someone diagnosed in BC and for which there was epidemiological support, or (2) those pediatric cases without WGS data but for which there was irrefutable epidemiological evidence of transmission. TB infections acquired outside BC were defined as those pediatric cases whose *Mtb* isolates had either a unique MIRU-VNTR pattern or who, by WGS, were >5 SNVs away from another isolate in BC and who had a documented history of residing in or traveling to a high-incidence TB country. Although other sources for these cases are possible—for example, a clinically diagnosed case unknown to the child, an adult source infecting a child before the study initiation, an unidentified visitor from outside BC, an individual whose active, infectious TB disease spontaneously resolved, or a case that left the province before diagnosis—these scenarios are considerably less likely. Pediatric cases not meeting either definition for place of acquisition were classified as having an unknown source.

5.2.6 Statistical analysis

Descriptive statistics were computed for demographic, clinical, and contact investigation data, both overall and stratified by birthplace. Unadjusted differences in characteristics between birthplaces were analyzed using the Fisher's exact test (categorical data) or the Kruskal-Wallis rank-sum test (non-normal continuous data). In addition, univariable analysis was conducted for factors related to locally acquired infection (yes/no) using the Chi-square test or Fisher's exact test, where appropriate. All analyses were executed in R (version 3.3.1).

5.3 Results

5.3.1 Demographics, clinical presentation, and epidemiology

From 2005 through 2014, a total of 98 children were diagnosed with active TB in British Columbia; 49 (50.0%) children had at least one culture-positive isolate available at the BCPHL (Figure 5-1). The median age of the study population was 14 (interquartile range [IQR]: 6–16); however, the age distribution varied significantly ($p = 0.023$) between CB and nCB children (Table 5-1, Figure 5-2), with nCB children almost always >10 years of age.

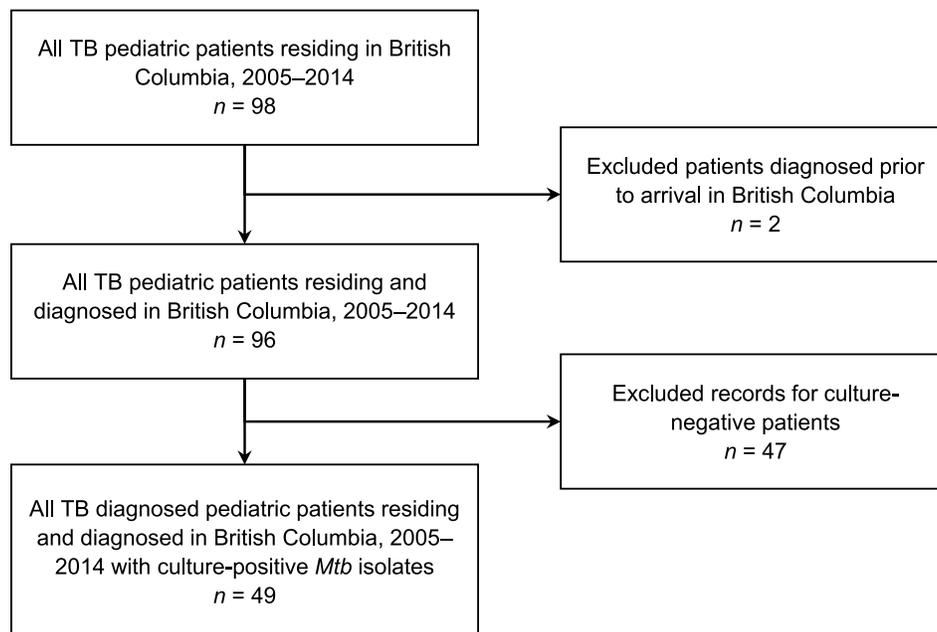


Figure 5-1. Pediatric TB study inclusion/exclusion criteria. Pediatric was defined as children <18 years of age.

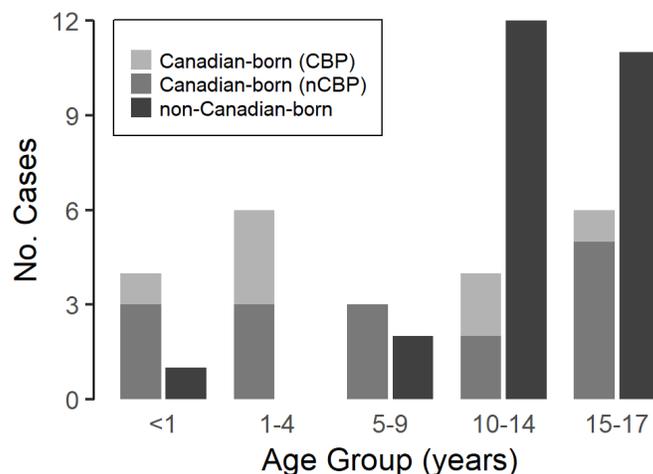


Figure 5-2. Pediatric age distribution by birthplace.

Abbreviations: Canadian-born parents (CBP); non-Canadian-born parents (nCBP).

Of the 26 children born outside Canada, 21 (80.8%) were born in Asia (including East, South-Eastern, and South-Central Asia) and the remaining five (19.2%) were born in Africa. All were born in high-incidence countries. With respect to ethnic community—defined as a combined measure of the child’s and/or parents’ region of birth—the highest proportion (42.9%) of the 49 pediatric cases were from South-Eastern Asia. Only 7 of 49 children (14.3%) came from a family that had resided in Canada for multiple generations.

Thirty-nine pediatric cases (79.6%) were diagnosed after symptomatic presentation to healthcare providers (**Table 5-1**). Six children (12.2%)—all Canadian-born (2 CBP, 4 nCBP)—were detected through contact investigations, and three (6.1%) nCB children were diagnosed as the result of immigration-related postlanding surveillance.

Clinically, 38 (77.6%) children had respiratory involvement (**Table 5-1**), and four children were characterized as having cavitary disease based on chest radiography (median age: 15.8 years, range: 14.7–17.9). One child in the study was HIV positive. Phenotypic drug susceptibilities revealed that 45 (91.8%) isolates were susceptible to all first-line antituberculous medications. Isoniazid monoresistance was seen in three individuals (6.1%)—all South-East Asian nCB adolescents with Indo-Oceanic lineage strains.

Table 5-1. Demographic and clinical characteristics of culture-positive pediatric TB cases. British Columbia, 2005–2014 (*n* = 49).^a

Characteristic	Overall	Canadian-born (CBP)	Canadian-born (nCBP)	Non-Canadian-born	<i>p</i> -value ^b
Totals	<i>n</i> = 49	<i>n</i> = 7	<i>n</i> = 16	<i>n</i> = 26	
Age, years					
Median (IQR)	14 (6–16)	4 (1–13)	7 (1–16)	15 (13–17)	0.023
Gender — <i>n</i> (%)					
Male	25 (51.0)	2 (8.0)	10 (40.0)	13 (52.0)	0.329
Female	24 (49.0)	5 (20.8)	6 (25.0)	13 (54.2)	
Ethnic community ^c — <i>n</i> (%)					
Multi-generational Canadian	7 (14.3)	7 (100.0)	–	–	–
South-Eastern Asia	21 (42.9)	–	7 (33.3)	14 (66.7)	
South-Central Asia	9 (18.4)	–	4 (44.4)	5 (55.6)	
East Asia	5 (10.2)	–	3 (60.0)	2 (40.0)	
Africa	7 (14.3)	–	2 (28.6)	5 (71.4)	
Disease Site — <i>n</i> (%)					
Respiratory	29 (59.2)	3 (10.3)	10 (34.5)	16 (55.2)	0.471
Non-Respiratory	11 (22.4)	1 (9.1)	3 (27.3)	7 (63.6)	
Respiratory + Non-Respiratory	9 (18.4)	3 (33.3)	3 (33.3)	3 (33.3)	
Respiratory ^d Smear — <i>n</i> (%)					
Positive	21 (53.8)	4 (19.0)	7 (33.3)	10 (47.6)	1.000
Cavitary					
Yes	4 (8.2)	1 (25.0)	1 (25.0)	2 (50.0)	0.620
No. Contacts					
Median (IQR)	5 (2–19)	17 (2–29)	5 (2–13)	5 (1–19)	0.819
Method of Detection — <i>n</i> (%)					
Symptoms	39 (79.6)	5 (12.8)	12 (30.8)	22 (56.4)	0.031
Contact Investigation	6 (12.2)	2 (33.3)	4 (66.7)	0 (0.0)	
Post-Landing Surveillance	3 (6.1)	–	–	3 (100.0)	
Incidental Finding	1 (2.0)	1 (100.0)	0 (0.0)	0 (0.0)	
Clustered ^e — <i>n</i> (%)					
Yes	24 (49.0)	7 (29.2)	7 (29.2)	10 (41.7)	0.011
No	25 (51.0)	0 (0.0)	9 (36.0)	16 (64.0)	

Abbreviations: CBP, Canadian-born parents; nCBP, non-Canadian-born parents; IQR, interquartile range.

^aPercentages have been rounded and may not total to 100%.

^bFisher’s exact test (categorical variables), and Kruskal-Wallis rank-sum test (non-normal continuous data).

^cEthnic community is derived from a combination of the region of birth for the pediatric case and parents of the child.

^dExcluded “other respiratory” sites e.g. pleura.

^eClustered = Yes where the isolate was identical by 24-locus MIRU-VNTR to another isolate in British Columbia (2005–2014).

The number of contacts varied considerably between individuals, ranging from 0 to 207, with 30 pediatric cases (61.2%) having fewer than 10 contacts (**Figure 5-3**). Extensive investigations (>50 contacts) were conducted around 3 acid-fast bacilli smear-positive children with respiratory TB. Contact investigation data suggested putative BC-resident source cases for 12 (24.5%) children; in four of these instances, the child was diagnosed before the adult source and served as the signal of an active infectious case in the community.

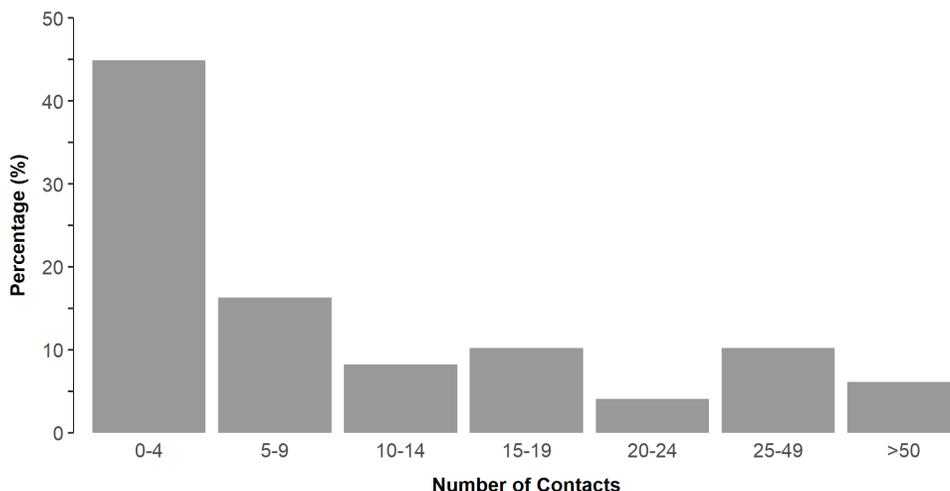


Figure 5-3. Number of contacts. Relative frequency representing the percentage of pediatric cases by number of individuals identified during contact investigation.

5.3.2 Molecular and genomic epidemiology investigation of putative sources

First, the 12 children with putative BC-resident sources identified through contact investigation were examined. In 11 instances, MIRU-VNTR and WGS supported the relationship between the child and assumed source. For the twelfth case, molecular data was not available for the adult TB contact; however, epidemiological evidence strongly corroborated the source of infection. For eight children (6 Canadian-born [nCBP], 1 Canadian-born [CBP], 1 non-Canadian-born), the source was an adult family member regularly residing in the same household as the child. For 4 children (3 Canadian-born [CBP], 1 Canadian-born [nCBP]), the source was a visitor to the household who resided elsewhere in BC or Canada. In one of these cases, contact investigation

had identified two plausible sources—a household member and a visitor; however, the combination of WGS results and epidemiological information suggested the visitor most likely transmitted to both the child and adult household member. One child did not meet either of the definitions for place of acquisition and was classified as having an unknown source.

5.3.3 Identification of infections acquired out of province

The MIRU-VNTR revealed that 25 *Mtb* isolates (51.0%) had a unique genotype, suggesting that the infection was likely acquired outside of the province. Indeed, 16 (64.0%) of the children with unique MIRU-VNTR patterns were born in high-burden countries.³³³ Of the nine Canadian-born children with unique MIRU-VNTR patterns, all had parents born outside Canada and seven had a confirmed travel history compatible with acquiring infection overseas. One of the 25 cases was ultimately determined to be the result of transmission in BC from a family member visiting from elsewhere in Canada, leaving 24 cases with an unknown source outside BC.

Also identified were 10 cases in which a pediatric *Mtb* isolate shared a MIRU-VNTR genotype with at least one other isolate from BC. In eight cases, the genomic distance between the pediatric case's isolate and the nearest BC isolate with an identical genotype precluded transmission (81–170 SNVs). In one case, the child's isolate was six SNVs away from an *Mtb* isolate from an adult born in the same country as the child and diagnosed in the previous year; however, there was no epidemiological link between the two, and a distance of six SNVs is thought to be incompatible with a transmission event occurring within a single year. In the absence of documented travel history, it was concluded that these nine children had been exposed to TB before emigration. The tenth case represented a non-Canadian-born sibling pair separated by a single SNV, for which epidemiological evidence pointed to infection before emigration from a common source.

Ultimately, whether through MIRU-VNTR or WGS, it was found that 33 (67.3%) culture-positive pediatric TB cases diagnosed in BC likely did not arise from local transmission (**Figure 5-4**). Epidemiological data suggested that most of these children acquired TB before their arrival in BC ($n = 23$), or through travel to their parents' birth country ($n = 8$). Two nCB children had

documented travel histories post arrival to Canada, making it unclear whether their infection was acquired before emigration or was travel-related.

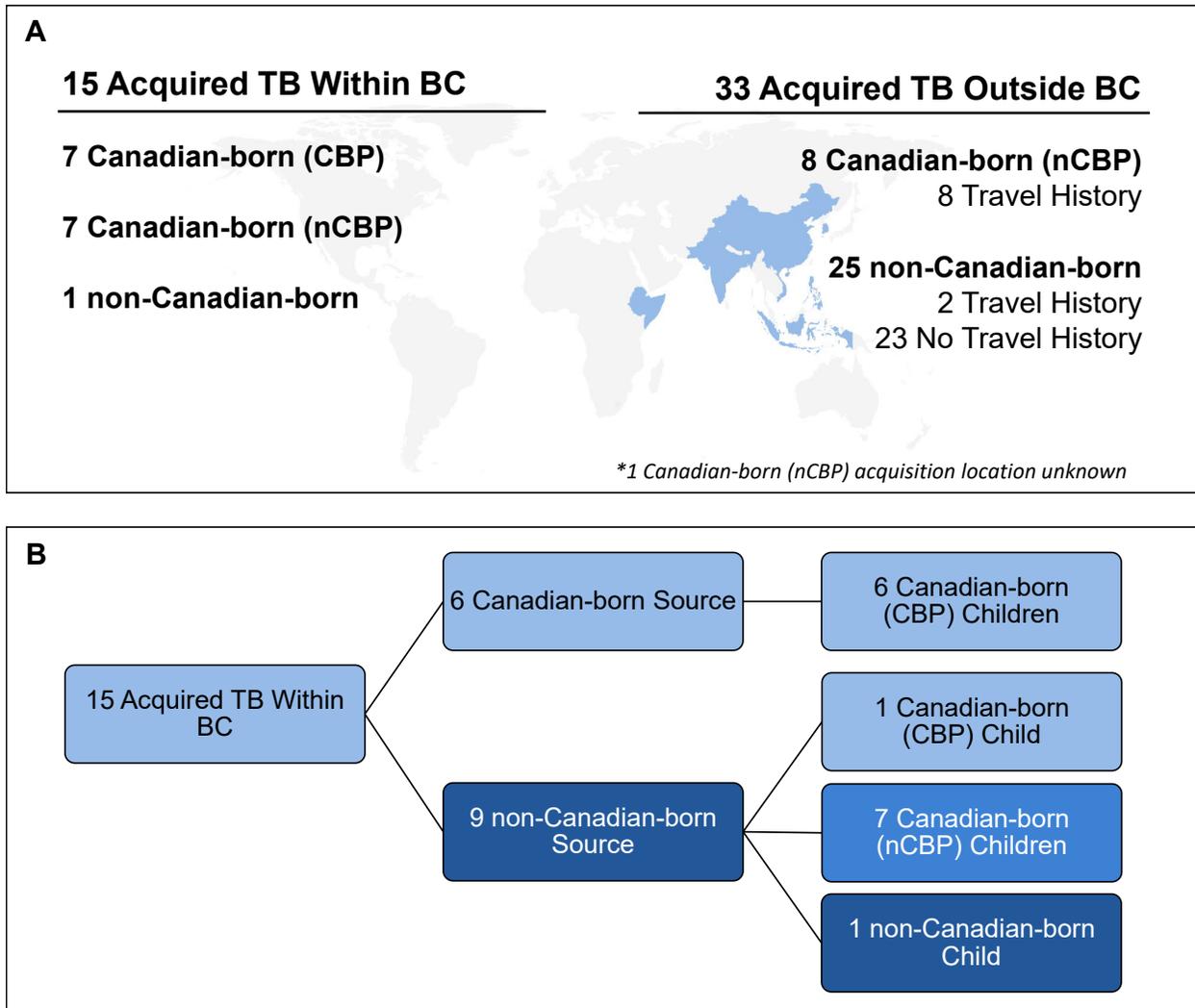


Figure 5-4. Pediatric tuberculosis investigation summary. Summary results of molecular epidemiological investigation of culture-positive pediatric tuberculosis (TB) cases, British Columbia (BC), 2005–2014. (A) Summarizes place of acquisition, and countries colored in blue correspond to travel history of individuals; (B) stratifies the birthplace of the pediatric case and source for those in which transmission occurred within BC. Canadian-born parents (CBP); non-Canadian-born parents (nCBP).

5.3.4 Identification of locally acquired infections

Fifteen (30.6%) culture-positive pediatric TB diagnoses in the study period resulted from presumed local transmission; this includes the 12 cases described earlier, for whom contact investigation suggested a BC resident as the likely source, and three additional transmissions identified through WGS, with ≤ 5 SNVs between the pediatric case and one or more adult cases (**Figure 5-5**). Of the 15 locally acquired cases, seven children were born in Canada to CBP. Only one WGS-confirmed source was a Canadian-born household family member; instead, three sources were Canadian-born visitors to the home, and in two cases, although a specific source was not identified, the children were infected with strains known to circulate within their communities. One Canadian-born child (CBP), most likely acquired TB from a nCB source (**Figure 5-5**) who was not identified through reverse contact investigation. Four of these children belonged to large, previously documented MIRU-VNTR clusters involving largely Canadian-born individuals: MClust-001 ($n = 56$), MClust-003 ($n = 39$), and MClust-055 ($n = 10$) (**Figure 5-5**).

Of the remaining eight cases—7 Canadian-born (nCBP) and 1 non-Canadian-born child—WGS (or epidemiology alone [$n = 1$]) suggested infection was acquired within BC from a nCB family member; seven regularly resided in the household. Two Canadian-born children (nCBP) belonged to large clusters comprising predominantly nCB individuals (MClust-011, MClust-187) (**Figure 5-5**).

The small sample size precluded multivariable regression; however, descriptive statistics (**Table 5-2**) indicate that local acquisition was associated with birth in Canada to Canadian parents, age under five, and infection with the Euro-American *Mtb* lineage.

Of the 49 cases, only one remained unclear at the conclusion of the investigation. The Canadian-born (nCBP) child had a unique MIRU-VNTR pattern suggesting acquisition overseas, but there was no documented travel history, nor did contact investigation suggest a putative source.

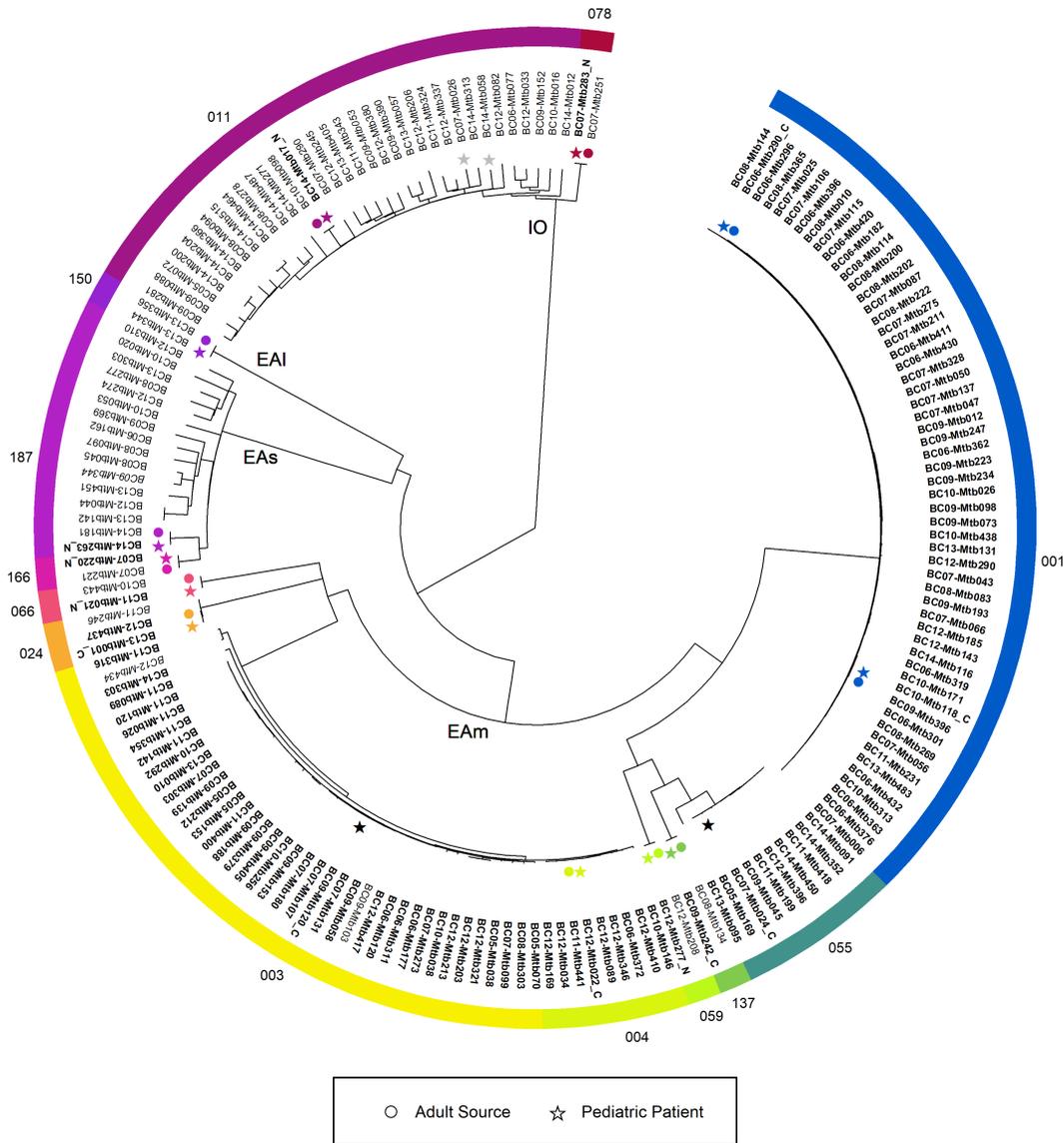


Figure 5-5. Pediatric analysis phylogenetic tree. Phylogenetic tree based on whole genome sequences of *Mycobacterium tuberculosis* isolated from all pediatric diagnoses resulting from WGS-confirmed transmission within BC ($n = 14$), and all adult isolates related by 24-locus MIRU-VNTR. Genotypic clusters are indicated by coloured bands. Pediatric (star) and adult (circle) cases are coloured where genomic epidemiology identified a clear source case ($n = 12$). Pediatric cases resulting from community transmission from an unknown source are indicated with black stars ($n = 2$). Grey stars indicate two pediatric cases who belong to a MIRU-VNTR cluster frequently seen in British Columbia (BC), for whom WGS indicated their infections were acquired outside BC. Bold tip labels indicate Canadian-born individuals; plain tip labels indicate non-Canadian-born individuals, and italicized tip labels indicate unknown birthplace. Canadian-born children with Canadian-born parents are annotated with “_C” and those with non-Canadian-born parents with “_N”. Internal branches are labelled by lineage: Euro-American (EAm), East-Asian (EAs), EAI (East-Asian Indian) and Indo-Oceanic (IO).

Table 5-2. Factors associated with locally acquired pediatric tuberculosis.
British Columbia, 2005–2014 ($n = 48$)^a.

Characteristic	Acquired Locally		<i>p</i> -value ^b
	Yes	No	
Birthplace			
Canadian-born (CBP)	7 (100.0)	0 (0.0)	<0.001
Canadian-born (nCBP)	7 (46.7)	8 (53.3)	
Non-Canadian-born	1 (3.8)	25 (96.2)	
Age, years			
<5	9 (81.8)	2 (18.2)	<0.001
≥5	6 (16.2)	31 (83.8)	
Gender			
Male	7 (28.0)	18 (72.0)	0.846
Female	8 (34.8)	15 (65.2)	
Travel History^c			
Yes	2 (16.7)	10 (83.3)	0.292
No	13 (36.1)	23 (63.9)	
Lineage			
Euro-American	10 (83.3)	2 (16.7)	<0.001
East-Asian	2 (25.0)	6 (75.0)	
East-African Indian	1 (10.0)	9 (90.0)	
Indo-Oceanic	2 (11.1)	16 (88.9)	

Abbreviations: CBP, Canadian-born parents; nCBP, non-Canadian-born parents.

^aSource unknown ($n = 1$).

^bChi-square test, (Fisher's exact test where appropriate).

^cTravel to a high-incidence TB country.

5.3.5 Household transmission of multidrug resistant tuberculosis

Multidrug resistance, defined as resistance to at least isoniazid and rifampin, was observed in one pediatric case with documented exposure to two adult TB cases, one with MDR-TB, and one with a pan-susceptible organism. The MIRU-VNTR genotyping placed this child and both adults into MClust-187; a cluster of cases ($n = 16$) involving an East-Asian lineage strain and predominantly nCB individuals with a median age of 66 (IQR: 47–87). Whole-genome sequencing analysis of MClust-187 revealed that the pediatric case was separated from the adult MDR-TB case by a single SNV; the adult contact with the pan-susceptible organism was 197 SNVs apart (**Figure 5-5**). Whole-genome sequencing revealed a second transmission pair in MClust-187—a household transmission between family members, both of whom harbored streptomycin-resistant organism—but neither individual was a pediatric case. With distances of 35–247 SNVs between them, the remaining 12 isolates in MClust-187 do not represent local transmission but rather a common region of birth.

5.4 Discussion

In the present study, genotyping and genomics was used to provide the first accurate estimate of TB transmission to children in a low-incidence setting, where the majority of all TB diagnoses (73.7%) are thought to represent reactivation of infections acquired abroad. By coupling a genotyping database including all culture-positive TB isolates diagnosed in British Columbia, Canada (2005–2014) to whole genome sequencing of clustered isolates and including epidemiological information, we find that one-third of culture-confirmed pediatric TB cases acquired their infection within BC. This rate is approximately three times that observed when genomics was used to interrogate transmission in a predominantly adult population in a similar low-incidence setting,⁷³ yet it is considerably lower than the pediatric transmission rate that would have been estimated based on MIRU-VNTR alone (49.0%).

A lack of laboratory testing and low diagnostic yields in children meant that only half of the notified pediatric cases had an isolate available for genotyping. This limits the present study somewhat, in that only transmission for cases with a culture-positive specimen can be reliably

assessed. Because novel technologies are making genome sequencing from primary specimens a reality,³⁷³ future studies of pediatric cases may be able to examine genomic data from a higher proportion of cases where specimens are submitted. The findings of this study indicate that to more fully understand TB transmission and the molecular epidemiology of a population, culture confirmation should be pursued in all cases.

The pediatric study TB population described here assorts into 3 distinct groups. Two thirds of the cases—largely nCB older adolescents—likely acquired their infection outside of BC, either before immigration or on a visit to their family’s country of origin. That this group tended to be older teenagers is in line with the findings of an American study reporting differing age distributions between American-born and immigrant children and attributing disease in the latter group to reactivation of latent TB infection (LTBI).³⁶⁸ The one third of TB cases likely to have acquired their infection in BC can be further divided into two groups: half of these cases were attributed to household transmission from a nCB family member, whereas the other half were community transmissions to Canadian-born children from Canadian-born adults, typically in the context of large outbreaks, two of which have been described previously.^{74,173} This latter group is notable—studies of pediatric TB cases in other Canadian provinces report 99–100% of children had nCB parents,^{87,365} but, in this BC-based study, 1 in 7 children diagnosed with TB had Canadian-born parents. This may reflect differing rates of local transmission between provinces; however, without WGS-based accurate estimates of transmission rates in each province, this hypothesis cannot be confirmed.

Although it is often stated that children with active TB serve as a sentinel case indicative of ongoing community transmission, this appears only to be true in particular sub-populations. Here, the only observed community transmission was in children of Canadian-born parents, where genomics confirmed that six of seven cases were attributed to community sources. No community transmission was observed in children with nCB parents, suggesting that extensive reverse contact investigations may not be warranted in this group. However, it should be noted that the study was limited to the detection of active TB infection as a marker of transmission and that not all transmissions result in active disease. Age <5 years is also often associated with local

transmission, and, although this was indeed true here, with most children <5 exposed via a household contact, over one third of locally acquired cases were in older children.

It was observed that 12 children had travel histories, and in at least eight of these cases the TB infection was likely acquired while on a trip to their parents' country of birth. Tuberculosis attributable to travel is often difficult to capture and separate from risk before immigration in nCB adults. Significant resources are dedicated to screening immigrants before and upon arrival in Canada; however, in subsequent encounters with the healthcare system, we do not reliably collect travel history for these individuals and tend to attribute their TB diagnoses as LTBI reactivation. The data indicate that travel to high-incidence settings to visit family poses an infection risk to children, thus it may also contribute to active TB cases among adults who travel to their country of birth. Immigrants traveling to visit friends and relatives in their country of origin are recognized as having increased risks for TB, particularly for long stays.³⁷³ It is interesting to note that at least three children in the study who likely acquired their infection during travel visited for less than the three-month indicator for screening recommended in the Canadian Tuberculosis Standards.⁶ Improved education around the risks of travel, better documentation of travel histories, and more aggressive screening protocols may be warranted in individuals returning from high-risk settings involving community-based travel. The findings of this study are: (1) in agreement with other studies regarding the risks of travel for children³⁶⁵ as well as adults^{374,375} and (2) suggest that new recommendations around screening individuals with community-based travel to high-incidence settings may be warranted.

The retrospective nature of this study meant that it was limited to epidemiological data recorded in BC's provincial TB registry and, in cases where genomics suggested a source that had not been named in the initial investigation, it was not possible to follow up these leads. Furthermore, the study was limited to use only molecular data to infer potential source cases diagnosed within the study window, because MIRU-VNTR and WGS data were unavailable for isolates obtained before 2005. This complicates source ascertainment for those children diagnosed in the first few years of the study; however, in each case, the available epidemiological data were sufficient to infer a reasonable source. Prospective WGS of all new culture-positive cases, recently

implemented by certain state and national mycobacterial reference laboratories, should allow for more timely and focused contact investigations, particularly in the context of larger outbreaks, where genotyping might suggest many possible sources, and in certain clusters involving nCB persons, where a genotypic relationship is infrequently borne out upon WGS.

Genomics is changing our understanding of TB transmission dynamics in low-incidence settings, and in the present study, its high resolution was used to more accurately estimate the proportion of pediatric TB attributable to local transmission. The study findings suggest that pediatric TB in BC is a mosaic and that factors including age, place of birth, and travel history must all be considered together when inferring a child's likely exposure. Thus, preventing future pediatric TB cases will likely require a flexible system with varying interventions, in some instances enhanced travel-associated screening, and in others, looking outside the home for source cases. Only through a combination of interventions will we be able to fully address this important issue.

Chapter 6: Whole Genome Sequencing for Improved Understanding of *Mycobacterium tuberculosis* Transmission in a Remote Circumpolar Region

6.1 Background

Canada's tuberculosis (TB) rate has been decreasing overall, yet rates remain elevated in particular populations and regions. Recent outbreaks in two areas of Canada's North—Nunavik and Nunavut—resulted in annual incidence rates higher than many low-income countries.^{6,82} However, this is not the case in all circumpolar settings, where public health efforts have contributed to declining TB rates. From 2006 through 2012, Yukon Territory (YT) reported a rate of 12.1 cases per 100,000 population. While this is over twice the national average of 4.8 cases/100,000; it is the lowest rate amongst Canada's Northern territories (25.4/100,000 in the Northwest Territories, immediately east of YT, and 194.3/100,000 in Nunavut).^{13,82} Alaska, located west of YT, has seen a sharp decrease in cases over the last few decades, reporting an average incidence of 8.1/100,000 (2006–2012), with most cases concentrated in rural communities—frequently inaccessible by road.^{82,376} Thus, while northern remote settings are often viewed similarly by population and public health programs, it is clear that with respect to TB, there are significant differences across these regions, likely explained by a combination of the robustness of regional public health, access to appropriate housing, geography, intra-community movement, and the populations themselves.³⁷⁷ Understanding the unique epidemiology of TB in each region is therefore vital to delivering tailored interventions to drive rates in circumpolar settings closer to the World Health Organization's elimination goals.

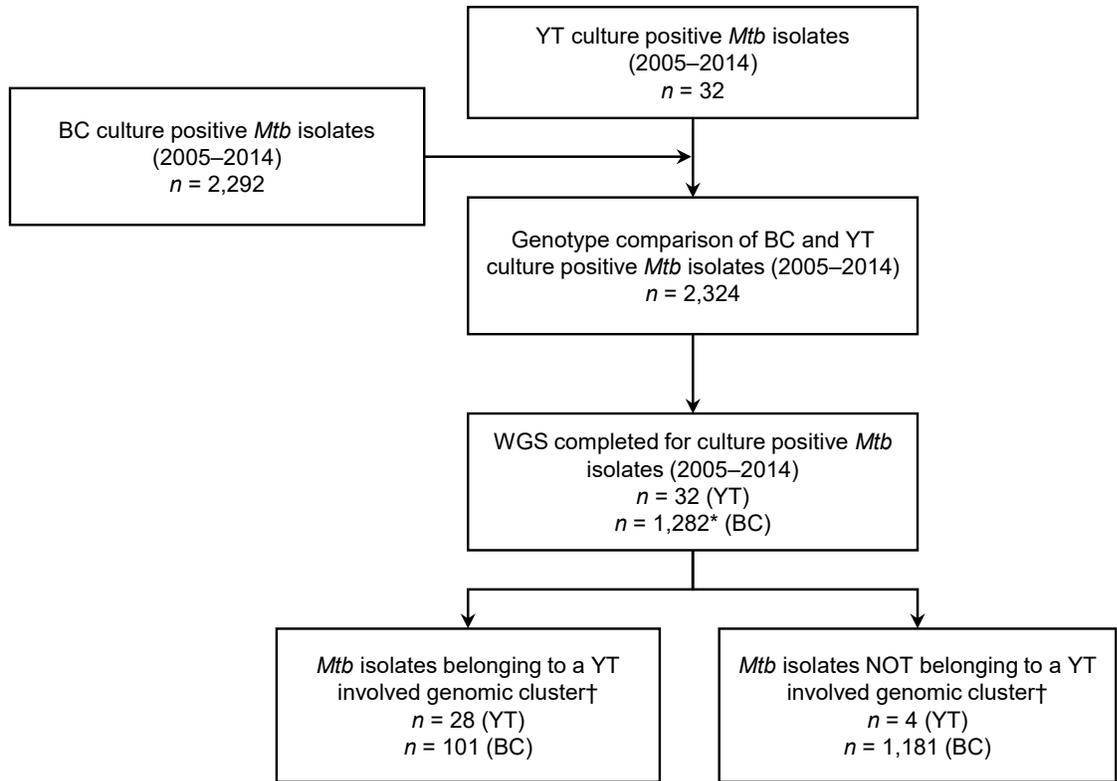
Genotyping programs have provided significant insights into the molecular epidemiology of TB in many low-incidence countries, helping to detect outbreaks,^{342,378} and more recently, genome sequencing has dramatically improved our understanding of both clustering and TB transmission in communities worldwide.^{74,117,149} However, only two studies to date have used this genomic epidemiology approach to examine transmission in remote northern locations: one in Nunavik, Québec¹⁶⁶—an Arctic region of Canada's North, and a second in Greenland, which used

genomics to detect “hotspot cases” responsible for chains of transmission.³⁷⁹ To better understand patterns of TB transmission in Yukon Territory, *Mycobacterium tuberculosis* (*Mtb*) genomes were sequenced from all culture-positive TB diagnoses in YT over a ten-year period—the first genomic epidemiology study of TB in this region. Recognizing that in contrast to many other northern regions in Canada, year-round highway access and multiple airports facilitate travel between YT and its southern neighbour, British Columbia (BC), YT *Mtb* genomes were also examined in the context of *Mtb* genomes sequenced in BC during the same time period. This unique cross-border comparison is possible because the BC Centre for Disease Control (BCCDC) and the BCCDC Public Health Laboratory (BCPHL) are contracted by YT to provide TB services such as laboratory diagnostics and case management support, and both jurisdictions access a shared data repository, thus allowing us to identify chains of transmission within and across YT/BC borders, and to fully describe the genomic epidemiology of tuberculosis in this remote circumpolar region.

6.2 Methods

6.2.1 Study setting and design

Yukon Territory is a sparsely populated (0.1 persons/km^2)³⁸⁰ territory located in the most northwestern region of Canada, immediately north of British Columbia. All tuberculosis cases diagnosed in YT are reported to the Yukon Communicable Disease Control (YCDC), and those in BC to the BCCDC. Care and treatment of individuals diagnosed with TB is the responsibility of YCDC, in partnership with Yukon Government Community Nursing and includes contact investigations (CIs) for newly diagnosed cases. The BCPHL receives all *Mtb* isolates for both YT and BC, and conducts routine diagnostic testing, universal 24-locus mycobacterial interspersed repetitive unit–variable-number tandem repeat (MIRU-VNTR) genotyping, and whole genome sequencing on request. The study population (**Figure 6-1**) included all YT culture-positive TB cases from 2005 through 2014 ($n = 32$), which were compared to TB cases diagnosed in BC during the same time period ($n = 2,292$), for which the BC study population has been previously described (**Chapter 3**).



*Included were two isolates for which genotyping results were unavailable.
 †Genomic cluster threshold of 5 SNVs

Figure 6-1. Study sample. Flow diagram summarizing the number of isolates from British Columbia (BC) and Yukon Territory (YT) belonging to a YT involved genomic cluster based on a five single nucleotide variants (SNVs) threshold.

6.2.2 Case-level information

Case-level clinical and demographic data, as well as epidemiological data collected during routine contact investigations (CIs), for all TB cases from BC and YT were extracted from the Integrated Public Health Information System (iPHIS). To classify community type for BC cases into metro (>190,000), urban/rural (40,001–190,000), rural (10,001–40,000), and remote (≤10,000) groups, the population density of the geographic service area in which each case resided was used. YT community types were classified by home postal code, with the second digit '1' in the forward sortation area indicating urban/rural, and a '0' indicating a remote community.

6.2.3 Laboratory methods

All *Mtb* isolates were obtained from specimens submitted to the BCPHL for routine testing. Isolates were revived from archived frozen stocks, DNA was extracted, and 24-locus MIRU-VNTR genotyping was performed as previously described (**Chapter 3**). Isolates lacking an amplicon peak at any locus were repeated with newly extracted DNA, and where there remained no peak at a single locus, the locus was coded as missing data and included in the analyses. All 32 culture-positive isolates of 38 notified cases in YT during the study period were successfully genotyped. These results were compared to genotypes of all culture-positive *Mtb* isolates from BC over the same period (**Chapter 3**). Whole genome sequencing (WGS) was completed for all 32 YT isolates as well as 1,284 BC isolates—including all those genotypically clustered by MIRU-VNTR to a YT isolate. WGS was completed using 125 bp paired-end reads on the Illumina HiSeq 2500 platform at Canada's Michael Smith Genome Sciences Centre (Vancouver, BC).

6.2.4 WGS analysis

The bioinformatics pipeline developed by Oxford University and Public Health England was used to analyze the resulting FASTQ files.²¹⁴ Reads were aligned to the *Mtb* H37Rv reference genome (GenBank ID: NC000962.2), and after masking for low complexity regions an average of 92% of the reference genome was covered. Single nucleotide variants (SNVs) were identified across all mapped non-repetitive sites. Genomic clusters were defined independently of MIRU-VNTR clusters and a unique identifier (WClustID) was assigned where isolates differed by ≤ 5 SNVs—a threshold reflecting recent local transmission.¹⁴⁹ Concatenated SNVs combined with epidemiological data collected through routine CIs and consultation with YCDC public health authorities were used to generate temporal transmission networks. Major lineage was predicted for each sequenced isolate based on lineage-defining SNVs.³⁷² FASTQ files for all genomes are available at NCBI under BioProject PRJNA413593 and PRJNA49659.

6.2.5 Statistical Analysis

Descriptive statistics were calculated for basic demographic and clinical information across two categories: (i) all cases diagnosed within YT, and (ii) BC cases with an *Mtb* isolate ≤ 5 SNVs to a YT case and thereby classified as “Related” (BC^R). Univariable analysis used the t-test for comparisons of mean age, and categorical variables were compared using Chi-square or Fisher’s exact test where appropriate. The frequency for which a MIRU-VNTR pattern was observed within the YT and/or BC^R populations was described, and to place MIRU-VNTR genotypes in the wider context of BC as a whole, genotypes were also compared to BC isolates not closely related to YT isolates based on genomic distance thresholds (>5 SNVs). These were classified as “Not Related” (BC^{NR}). A dendrogram based on 24-locus MIRU-VNTR genotyping patterns was generated using the categorical (Hamming) distance and UPGMA (unweighted pair group method with arithmetic mean) algorithm. All statistical analyses were done in R v3.4.1.

6.3 Results

6.3.1 MIRU-VNTR and WGS provide different estimates of clustering

From 2005 through 2014, 32 individuals were diagnosed with culture-positive TB in Yukon Territory. MIRU-VNTR genotyping grouped 21 of these cases into three clusters (3–13 YT isolates/cluster), yielding a clustered proportion of 65.6% within the territory. One YT isolate had an untypable locus yet matched a cluster unique to YT for the other 23 typable loci. Six YT isolates had MIRU-VNTR patterns that were unique amongst the YT population yet clustered with isolates in BC, bringing the total number of MIRU-VNTR clusters across both jurisdictions containing at least one YT case to nine (**Figure 6-2**). Four YT isolates remained unclustered after comparison with all BC isolates; however, all four were within one or two loci of a YT and/or BC genotype cluster.

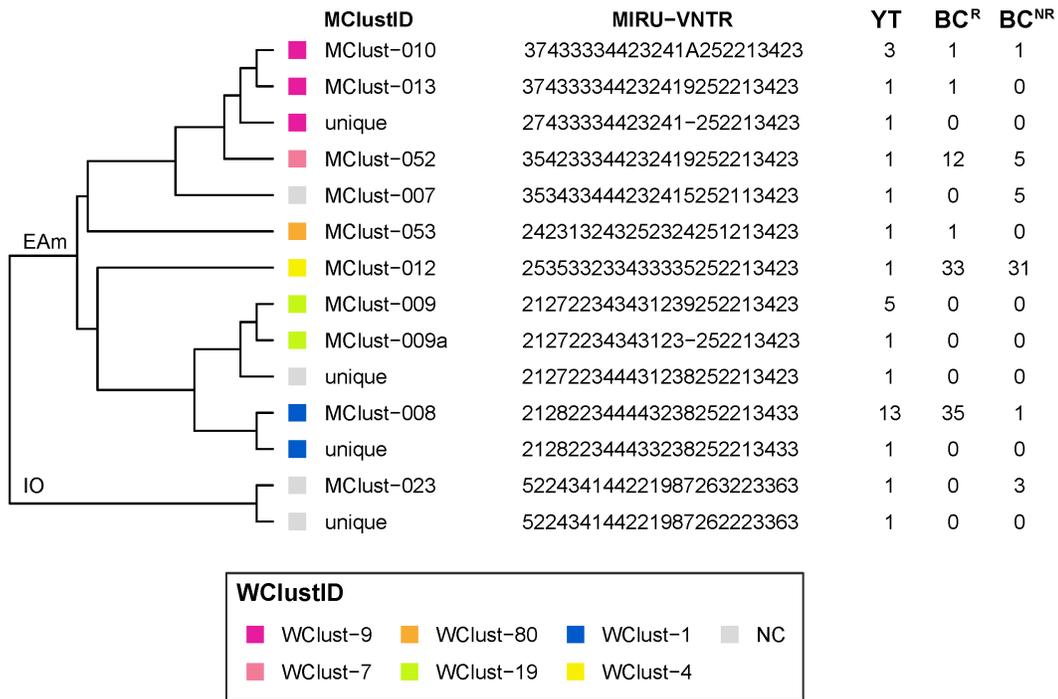


Figure 6-2. Population structure of *Mycobacterium tuberculosis* in Yukon Territory. Dendrogram based on 24-locus MIRU-VNTR genotypes of *Mycobacterium tuberculosis* isolates collected in Yukon Territory (YT) from 2005 through 2014. MIRU-VNTR clusters (≥ 2 isolates) were assigned a unique MClustID. The number of times each MIRU-VNTR pattern was observed in the YT, BC^R—(within 0–5 SNVs of a YT isolate), and BC^{NR}—BC (>5 SNVs to a YT isolate) populations are indicated in the right-hand side columns. Coloured squares represent the whole genome sequencing cluster, and isolates >5 SNVs from isolates in YT or British Columbia (BC) were considered not genomically clustered (NC). WGS clusters (≥ 2 isolates within 0–5 SNVs) were assigned a unique WClustID independent of MIRU-VNTR. Lineage is indicated at the root. Abbreviations: EAm, Euro-American, IO, Indo-Oceanic. Order of loci: MIRU 04, MIRU 26, MIRU 40, MIRU 10, MIRU 16, MIRU 31, 424, 577, 2165, 2401, 3690, 4156, 2163, 1955, 4052, MIRU 02, MIRU 23, MIRU 39, MIRU 20, MIRU 24, MIRU 27, 2347, 2461, 3171.

Genomics provided a higher resolution view of clusters suggestive of recent transmission, merging several MIRU-VNTR clusters that differed by a single locus or had an untypable locus into single groups supported by contact investigation data, and in other cases revealing that MIRU-VNTR clustered isolates, such as those belonging to MClust-023, were not truly clustered in a way that would suggest recent local transmission (**Figure 6-2**). Using a five SNV threshold, six genomic clusters were identified with at least one YT case, involving a total of 28 YT and 101 BC^R isolates and ranging from two to 59 isolates (**Figure 6-3**). Another YT isolate was within 20 SNVs of a genomic cluster, while the remaining three isolates were >200 SNVs away from any other YT isolate. By WGS, the clustered proportion was 28/32 (87.5%) when YT isolates were considered alongside BC isolates, and 25/32 (78.1%) considering only isolates among YT residents. With the exception of two Indo-Oceanic lineage isolates, all other YT isolates (94.1%) belonged to the Euro-American lineage.

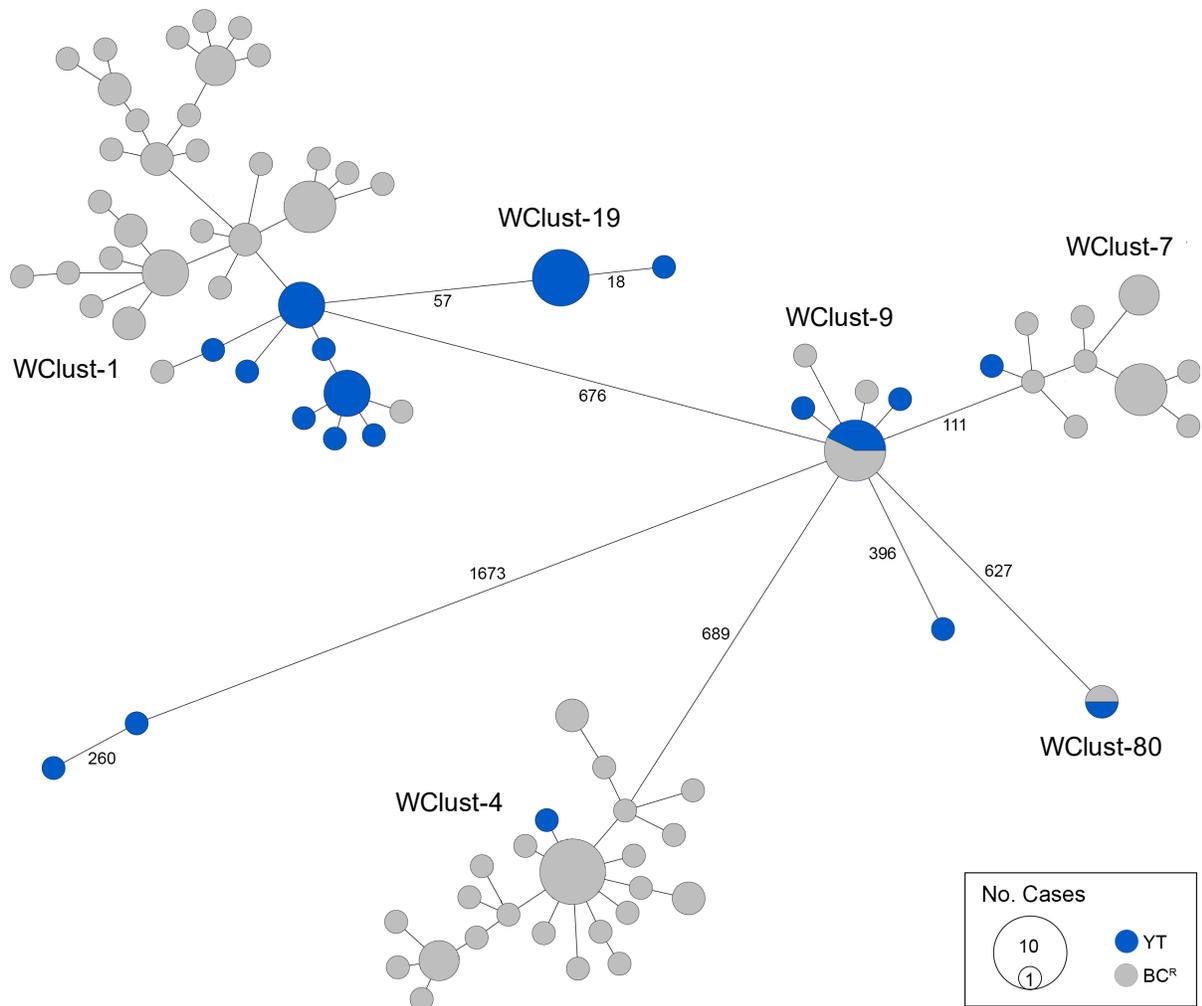


Figure 6-3. Yukon Territory *Mycobacterium tuberculosis* isolates in the context of related BC isolates. Minimum-spanning tree based on whole genome sequences of *Mycobacterium tuberculosis* (*Mtb*) isolates from the Yukon Territory (YT), Canada study population ($n = 32$) and closely related (five single nucleotide variants [SNVs] threshold) isolates from British Columbia (BC) ($n = 101$). The size of each circle is proportional to the number of isolates, and circles are coloured in blue to represent the YT study population and grey for the BC population. Unique cluster identifiers (WClustID) are indicated for isolates in genomic clusters. The number of SNVs between isolates with >5 -SNVs are indicated along the connecting branches.

6.3.2 Genomically related cases across jurisdictions are similar clinically

Comparing all YT cases to the genomically related BC^R cases ($n = 101$), characteristics were found to be similar across both populations, including the mean age of 45.8 years (standard deviation [SD] ± 16.7) and 46.8 years ($SD \pm 11.9$) for YT and BC^R individuals, respectively. Both groups were predominantly Canadian-born, with 93.8% of the YT study population and 88.9% of BC^R persons born in Canada (**Table 6-1**). The proportion of individuals with a clinical presentation associated with TB transmission was high in the YT and BC^R populations, with respiratory TB diagnosed in 90.6% of YT and 89.1% of BC^R individuals. Likewise, the smear-positive TB proportion was high—greater than 82% in YT and BC^R persons. Of note, the proportion of individuals with cavitary TB was over 1.5 \times higher in the YT population compared to BC^R individuals, with cavitary disease in 37.5% (12/29) of YT persons ($p = 0.099$). With respect to risk factors for transmission,³⁸¹ the majority of individuals (YT: 71.9%, BC^R: 61.5%) reported ≥ 1 risk factor (HIV, illicit drug use or alcohol misuse). Reflecting the differing demographics between the two settings, the majority of YT individuals resided in remote (84.4%) regions, compared to those in BC^R where the majority resided in metro areas (82.2%).

Table 6-1. Characteristics of Yukon study population. Demographic and clinical characteristics of culture positive cases across Yukon and genomically related cases in British Columbia, Canada, 2005–2014.^a

Characteristic	No. Cases (%)		p-value ^b
	YT	BC ^R	
Totals	<i>n</i> = 32	<i>n</i> = 101	
Age, years			
0–24	3 (9.4)	1 (1.0)	0.086
25–44	10 (31.2)	40 (39.6)	
45–64	14 (43.8)	50 (49.5)	
65+	5 (15.6)	10 (9.9)	
Gender			
Male	21 (65.6)	69 (68.3)	0.777
Community			
Metro	0 (0.0)	83 (82.2)	<0.001
Urban/Rural	5 (15.6)	7 (6.9)	
Rural	0 (0.0)	8 (7.9)	
Remote	27 (84.4)	3 (3.0)	
Birthplace ^c			
Canada	30 (93.8)	88 (88.9)	0.734
Disease Site			
Respiratory	29 (90.6)	82 (81.2)	0.344
Non-Respiratory	3 (9.4)	11 (10.9)	
Respiratory + Non-Respiratory	0 (0.0)	8 (7.9)	
Respiratory ^d smear			
Positive	23 (82.1)	76 (83.5)	1.000
Cavitary disease			
Yes	12 (37.5)	23 (22.8)	0.099
Risk Factors ^e			
None	9 (28.1)	30 (38.0)	0.325
≥1	23 (71.9)	49 (62.0)	

Abbreviations: BC^R, British Columbia Related (*Mycobacterium tuberculosis* isolates ≤5 SNVs to YT study population); SNVs, single nucleotide variants; YT, Yukon Territory.

^aPercentages have been rounded and may not total to 100%.

^bChi-square test, (Fisher’s exact test where appropriate).

^cData unavailable *n* = 2 (BC^R).

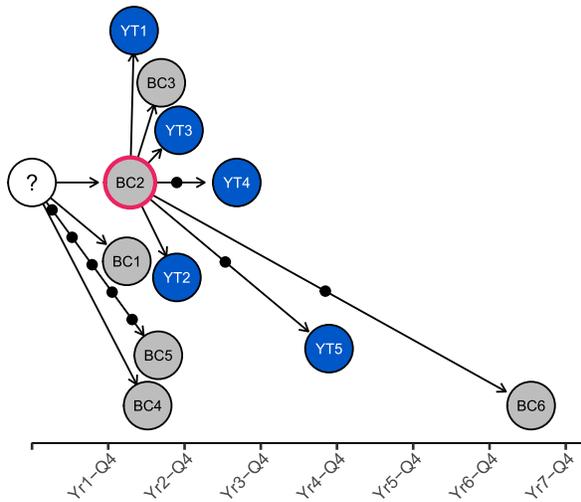
^dExcluded “other respiratory” sites e.g. pleura.

^eRisk Factors = HIV, illicit drug use, or alcohol misuse; data unavailable for 1 or more risk factor in BC^R population (*n* = 22).

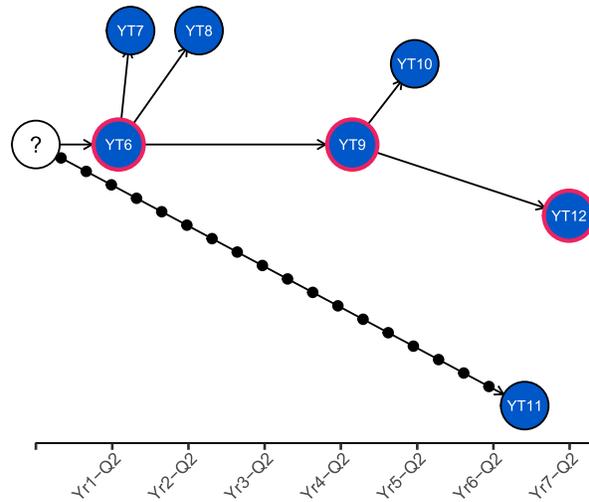
6.3.3 Transmission reconstruction

To characterize person-to-person spread of TB within YT, temporal transmission networks were constructed using WGS results combined with epidemiological data for the three genomic clusters with sustained transmission between YT persons—WClust-1, WClust-9, and WClust-19 (**Figure 6-4**). Although *Mtb* isolate YT13 is above the five SNV threshold set for recent transmission, it is within 18 SNVs of WClust-19—a cluster genotypically and genomically unique to the YT population—and was therefore included in the reconstruction figure. This case likely represents reactivation of a previously acquired infection with a strain circulating within YT. For WClust-1, a large cluster with discrete minimum spanning tree branches in both YT and BC, only the branch of YT isolates was included, together with the two closely related BC isolates (**Figure 6-3**).

WClust-9



WClust-19



WClust-1

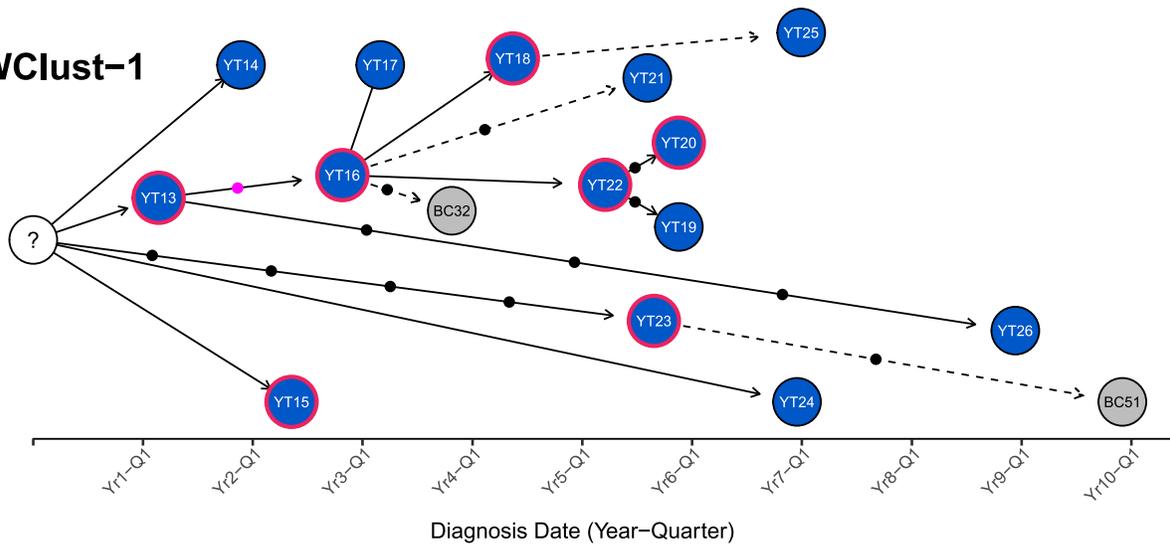


Figure 6-4. Yukon WGS transmission reconstructions. Transmission networks of three *Mycobacterium tuberculosis* genomic clusters (based on a five single nucleotide variants [SNVs] threshold) representing transmission in Yukon Territory (YT), Canada (2005–2014). Blue circles represent YT isolates and grey British Columbia (BC) isolates. A red outline around a circle represents an individual who is acid fast bacillus smear-positive with cavitory TB disease. Solid lines indicate strong epidemiological linkages, and dashed lines indicate weak epidemiological linkages. SNVs acquired over time are represented by dots between isolates. The pink dot represents the presence of a minority-variant which becomes fixed in all isolates in the subsequent transmission chain.

Each of the three clusters differ slightly. WClust-19 is the only cluster exclusively comprising YT individuals, whereas WClust-1 and WClust-9 had one or more BC persons with related isolates. Within WClust-1 the BC cases may have acquired TB from a YT individual, whereas in WClust-9 a BC individual likely transmitted TB to a number of BC and YT cases. SNV distances ranged within clusters; however, WClust-19 saw no genomic variation in the transmission chain stemming from YT8, despite up to six years between disease acquisition and diagnosis. WClust-9 has four BC isolates 0–5 SNVs from those in YT (**Figure 6-4**). However, with the exception of BC2 there are no known epidemiological connections between these cases that would suggest a common source not identified through CIs. Interestingly, all three transmission clusters each have at least one individual acting as the source of three or more active culture-positive cases.

WClust-1 represents the largest YT cluster. CIs revealed that many of the individuals were social contacts of one another, with at least two individuals suspected of giving rise to multiple secondary cases. Here, genomics identified a minority variant (at the SNV site, 15% of reads had adenine [A] and 85% were cytosine [C]) in the sample from YT18, whereas in samples from subsequent cases, the minority SNV was fully fixed, confirming this individual as the most likely source for the cases that followed (**Figure 6-5**). Genomic data also confirmed the inclusion of three Yukon (YT23, YT25 and YT27) and two BC isolates (BC31 and BC49) in this cluster, despite no apparent epidemiological linkages to each other or other cluster members.

While each of the three genomic clusters had unique features, all had at least one individual source of multiple culture-positive secondary cases, and all spanned several years, with some individuals progressing rapidly to active disease, and others reactivating after a long period of latency.

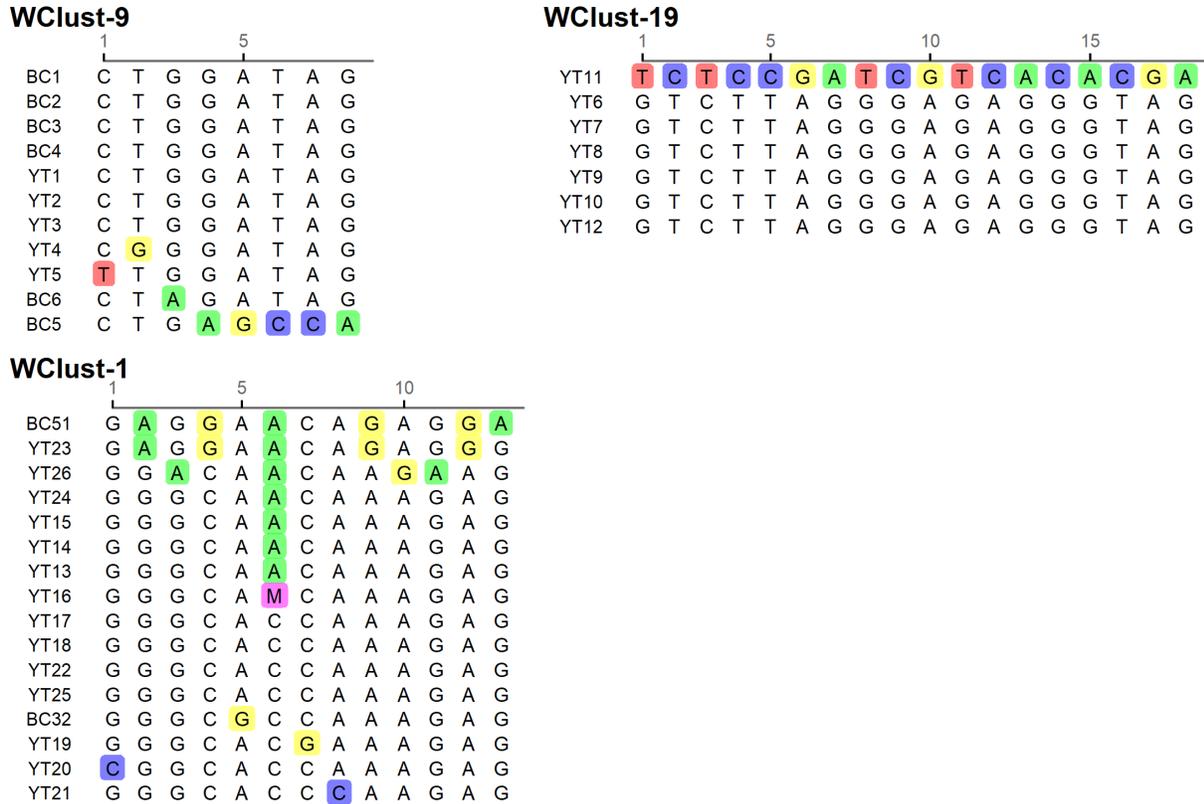


Figure 6-5. Transmission clusters—SNV alignments. Multiple sequence alignments of concatenated single nucleotide variants (SNVs) for three *Mycobacterium tuberculosis* genomic clusters (based on a five SNVs threshold) representing sustained transmission in Yukon Territory, Canada (2005–2014).

6.4 Discussion

The genomic epidemiology of tuberculosis in northwestern Canada over a ten-year period was described, finding that persons diagnosed with TB were largely Canadian-born with nearly all cases attributable to local transmission, consistent with the epidemiology of TB elsewhere in Canada’s North.^{6,166}

Genomic data, combined with detailed epidemiological data, allowed for the reconstruction of likely transmission routes among the three large clusters. It was found that, as is true for a number of infectious diseases, a small number of individuals account for a disproportionate number of secondary cases—the phenomenon of “super-spreaders”.³⁸² Understanding the risk

factors and epidemiological characteristics driving super-spreading in a community is important for better prioritizing TB prevention and care programs. In the YT study population, the proportion of individuals with clinical risk factors frequently associated with transmission, such as cavitory disease and smear positivity,³⁸¹ was quite high, and anecdotal evidence from the local public health team suggested that delays in diagnosis might have also contributed to transmission. A recent publication⁴ discussed the various drivers of TB transmission outside clinical risk factors, including diagnostic delays, which increase the potential for disease progression and transmission,^{54,383} particularly amongst highly mobile, socially connected, and infectious individuals.

Given the shared border between Yukon Territory and BC, transmission across jurisdictions was also examined. Including genomically related BC isolates increased the estimate of clustering for YT isolates, suggesting that estimates derived from provincial or territory data alone likely underestimate transmission rates in relation to remote settings. Cross-border transmission appears to occur in both directions—in several cases YT residents likely transmitted to BC residents via social/community connections with YT residents reporting travel/residential histories in both northern BC communities and larger metropolitan regions. Additionally, three YT cases had isolates that clustered only with BC isolates and likely acquired their infections within BC, while a BC source was linked to six YT cases in WClust-9.

Given the low genomic variation between cases, with most cases differing by 0–1 SNVs, the cluster reconstructions were only possible thanks to the detailed epidemiological information collected by the local public health team. Such minimal variation across multiple hosts over many years is not uncommon, and has been previously described in outbreaks elsewhere in Canada.¹¹⁷ This observation reinforces the need for comprehensive contact investigation data coupled to genomics to fully understand regional epidemiology, though it is important to note that because genomic studies currently require *Mtb* culture, culture-negative TB cases are excluded from reconstructions. These cases are less likely to contribute to transmission due to low bacterial loads but cannot be completely excluded. TB diagnoses prior to the study period are also not captured here.

Understanding TB transmission dynamics is key to the design and delivery of effective evidence-based interventions to prevent the continuing spread of TB. Here, it was found that WGS combined with detailed CIs information allowed for a more refined picture of transmission than either method alone, or with the use of MIRU-VNTR. The implementation of routine genotyping and WGS is recommended with linkage of these results to epidemiological and CIs data. Knowledge of the genomic connections between isolates in YT and BC will support TB programs and improve communication and understanding of transmission across jurisdictional boundaries.

Chapter 7: Comparison of Traditional Field Epidemiology and Whole Genome Sequencing to Understand Tuberculosis Transmission in a Remote Setting

7.1 Background

Tuberculosis (TB) remains an important public health concern in Canada, particularly in northern rural and remote areas where endemic spread of TB is commonplace.¹⁶⁶ Understanding patterns of transmission in these settings is an integral part of developing evidence-based prevention and care strategies and prioritizing public health resources—this includes understanding the burden of disease resulting from recent local transmission versus reactivation of historic latent TB infection (LTBI), as well as understanding the nature of recent transmission. This latter point is critical for improving TB services in a region—understanding the clinical, demographic, and/or epidemiological factors driving TB transmission is vital to developing informed prevention programs, screening activities, and contact investigations (CIs) and ultimately preventing the continued spread of TB.

Field-based epidemiological investigation is used to identify both infected contacts, secondary active cases, and possible sources of a given case, and for decades was the only means to detect transmission.³⁸⁴ In recent years, a combination of field and molecular epidemiology has been used in many settings—contact data collected through interviews of recently diagnosed individuals may reveal the potential links between cases, while genotyping techniques identify related *Mycobacterium tuberculosis* (*Mtb*) isolates and can confirm or refute a potential transmission event. Now, several studies have shown that whole genome sequencing (WGS) yields more accurate transmission reconstructions than the approaches based on genotypic data.^{74,149,158} In British Columbia (BC), Canada, WGS was used retrospectively to better understand a large TB outbreak⁷⁴ and in real-time as part of the management of a second large outbreak in BC,^{173,174} and it has now become routine practice in the UK to use WGS to identify clusters of related cases for public health follow-up.

Despite global interest in WGS as a tool for understanding TB epidemiology and a continuously expanding dataset of publicly available *Mtb* genomes, there are gaps in our understanding of how useful this new technique is. There are technical questions around how consistent *Mtb* mutation rates are, particularly during latent infection versus active disease^{385,386} and from human host to human host,^{147,300} as well as around how to identify transmission-informative variants in the many repetitive elements within the *Mtb* genome.^{387,388} There are also questions surrounding its utility. In well-resourced rural and remote settings, where detailed contact tracing and interview data are often available for each case, it is not known whether WGS offers any benefit over the current standard of care—interpreting genotyping data in the context of this rich field epidemiological data—and there have only been limited comparisons of how useful the molecular data alone is, whether genotypic or genomic.^{166,389} Furthermore, there has been no qualitative feedback data from frontline public health personnel describing if/how molecular data improved their ability to understand a cluster of cases in a remote setting.

The Yukon Territory (YT), located in Canada’s northwest, has a higher TB incidence (12.1 per 100,000) than the Canadian average (4.9 per 100,000), yet lower than other northern Canadian settings.^{82,84} As described in **Chapter 6**, the majority of YT residents diagnosed with TB are Canadian-born (93.8%) and live in remote regions (84.4%). All YT TB cases are managed by a small team of public health professionals, many of whom have deep and long-standing ties to the territory; this strong tradition of engagement between public health, community nurses, and YT’s communities means the local public health unit has uniquely detailed insights into the social networks underlying YT’s TB clusters. This, coupled to a small, remote population with frequent travel to BC, yet little in- or out-migration, makes this an ideal population in which to explore the utility of genomic data in enhancing contact investigations. Here, the utility of whole genome sequencing versus 24-locus mycobacterial interspersed repetitive unit–variable-number tandem repeat (MIRU-VNTR) coupled to robust traditional field epidemiology for identifying transmission, outbreaks, and reactivation of LTBI was determined. Two independent teams—one working with MIRU-VNTR and detailed contact investigation data and the other working with WGS data and basic clinical and epidemiological information—each reconstructed the most likely transmission pathways for every culture-positive TB case diagnosed in YT from 2005–

2014. The teams then met to jointly infer the most plausible transmission network, given both the social contact investigation and WGS data. Here, the results of these reconstructions are presented, as well as qualitative user feedback on the utility of genotyping and genomics in a remote setting with a comprehensive TB contact investigation program.

7.2 Methods

7.2.1 Study setting and design

The study took place in Yukon Territory (YT), Canada, a remote arctic/sub-arctic territory with a population of approximately 38,400 in 2017, spread over an area of more than 470,000 square kilometers.^{380,390} The study population included all 32 persons diagnosed in YT with culture-confirmed TB from 2005 through 2014 (84.2% of all 38 diagnoses). Yukon Communicable Disease Control (YCDC), in partnership with Yukon Government Community Nursing, is responsible for patient care and treatment, with contracted TB services including laboratory diagnostics, case management support, and access to a shared data system provided by the British Columbia Centre for Disease Control (BCCDC) and the BCCDC Public Health Laboratory (BCPHL).

7.2.2 Bacterial culture, genotyping and whole genome sequencing

All *Mycobacterium tuberculosis* (*Mtb*) isolates were obtained from specimens submitted to BCPHL for routine clinical testing of tuberculosis. *Mtb* isolates were cultured, DNA extracted, and 24-locus MIRU-VNTR genotyping was carried out using standard methods.⁷⁷ All samples were sequenced on the Illumina HiSeq 2500 platform (Illumina, San Diego, USA) at the Michael Smith British Columbia Genome Sciences Centre (Vancouver, Canada) to produce 125 bp paired-end reads, which were mapped to the H37Rv reference genome (GenBank ID: NC000962.2) using the Public Health England/Oxford University bioinformatics pipeline.²⁶⁷

7.2.3 Source identification by field and molecular epidemiology

The first team (field-based) comprised YCDC nursing staff and program managers responsible for treatment and care of all persons diagnosed with TB and their contacts in the territory. They reviewed detailed notes from CIs for each individual in the study, and were provided with the MIRU-VNTR cluster each isolate belonged to, along with a general description of these clusters across BC and YT (e.g. size, geographic distribution, basic demographics) (**Figure 7-1**). The team was provided with a structured spreadsheet and was asked to identify each case's most likely source from the following options: a specific individual within YT; an unknown individual within YT; acquisition from an unknown individual through travel outside YT; or reactivation of LTBI acquired prior to the study period. Deliberations took into consideration MIRU-VNTR data, along with each TB case's prior contact history, symptom onset date, tuberculin skin test (TST) records, and transmission risk factors including acid-fast bacillus (AFB) smear status and presence of cavitory disease. Respondents were also asked to provide a confidence score to each presumed source: 0 – not at all confident, 1 – somewhat confident, 2 – very confident, 3 – certain.

MClust-008
<p>This cluster belongs to the Euro-American lineage, and the 24-locus MIRU-VNTR pattern has been seen 40 times in British Columbia from 2005 through 2014 in the following BC health service delivery area(s): South Vancouver Island, Vancouver, Fraser East, Fraser North, Fraser South, Northwest</p> <p>The overall demographics of this MIRU-VNTR cluster (BC & YT): 62% male, median age (years) = 45 (IQR: 35–55) and 95% Canadian-born.</p>

Figure 7-1. MIRU-VNTR cluster summary example.

7.2.4 Source identification by genomic epidemiology

The second team (genomic-based), comprising TB genomics experts from BCCDC, had access to WGS data for each YT isolate (see **Chapter 6**), as well as WGS data from all MIRU-VNTR clustered *Mtb* isolates from cases diagnosed in BC as part of the ten-year retrospective study (**Chapter 3**), including those that matched at least 23/24 MIRU-VNTR loci with a YT isolate. Genomic clusters were defined using a threshold of five single nucleotide variants (SNVs)¹⁴⁹ and were assigned a unique identifier (WClustID). Using WGS data but no field epidemiological information, this team independently constructed putative transmission networks from the genome sequences of all YT study isolates ($n = 32$) and any BC isolates within five SNVs of a YT isolate ($n = 101$). A minimum-spanning tree (MST) was generated and coloured by MIRU-VNTR cluster ID (MClustID), with labels indicating the genomic WClustID. The team subsequently refined each network with basic epidemiological data, including diagnosis date, place of residence, acid-fast bacilli smear status, chest radiology results and risk factors (HIV, illicit drug use, alcohol misuse) from the integrated Public Health Information System (iPHIS), but did not have access to any contact investigation or social network information. The team identified each case's most likely source from the above options. A confidence score was assigned to each source identified as described above.

7.2.5 Source identification consensus

At a joint, in-person meeting with both teams, each YT study case was reviewed and a consensus reached regarding the most plausible source, given the combination of WGS and field epidemiological data. During this meeting, informal training and background information regarding the interpretation and limitations of genotyping and WGS data were provided through a PowerPoint presentation and discussion of the genotyping and genomic data for each case.

7.2.6 Qualitative assessment

To examine the YCDC's team knowledge, attitudes, and practices around genotyping and genomic services, an online, multiple-choice survey was conducted both before and after the in-person consensus meeting (see **Appendix Table A-1**, **Appendix Table A-2** for questions). At the conclusion of the consensus meeting, a semi-structured group interview was conducted with the TB prevention and care team who completed the field epidemiology-based source identification—three nurses, a program manager (also a nurse) and Yukon's Chief Medical Officer of Health. The interview's objective was to collect qualitative feedback on the usefulness of molecular and genomic data for investigation of TB cases, the potential added value of MIRU-VNTR and WGS, and how this information could be used prospectively (see **Appendix Table A-3** for questions). The interview questions served as prompts to structure the conversation, but all persons were free to comment, at any depth. The interview was recorded, and manually reviewed. Using a thematic analysis method,³⁹¹ statements were coded and categorized according to identified common themes.

7.2.7 Statistical methods

The insights into TB transmission provided by field epidemiology/genotyping and WGS were compared by analyzing the outcomes of each investigation at three levels of resolution—individual (i.e. did the source identified by either team match the source determined at the consensus meeting), population-level (i.e. was the case ascribed to the correct transmission cluster—defined genomically) and probable location of TB acquisition (YT, BC, other province/territory, or outside Canada). All statistical analyses were completed using R (v3.4.1). Agreement between results for identified source from the field- and genomic-based investigations was measured using Cohen's kappa. Kappa values of <0.2, 0.21–0.40, 0.41–0.60, 0.61–0.80, 0.81–1.00 indicate poor, fair, moderate, good, and very good agreement, respectively.³⁹² Fisher's exact test was used for comparisons of proportions. Correlations between qualitative variables—level of certainty assigned to source identification, were assessed using Spearman's rho.

7.3 Results

Detailed clinical and epidemiological information, social contact data, 24-locus MIRU-VNTR genotypes, and whole genome sequences were available for all 32 (100%) of the YT study isolates diagnosed from 2005 through 2014. Typically, 1–2 cases were diagnosed each quarter, with the epidemiological curve (**Figure 7-2**) showing a notable peak corresponding to an increase in cases matched to one of the three circulating YT-specific *Mtb* strains, as defined by WGS.

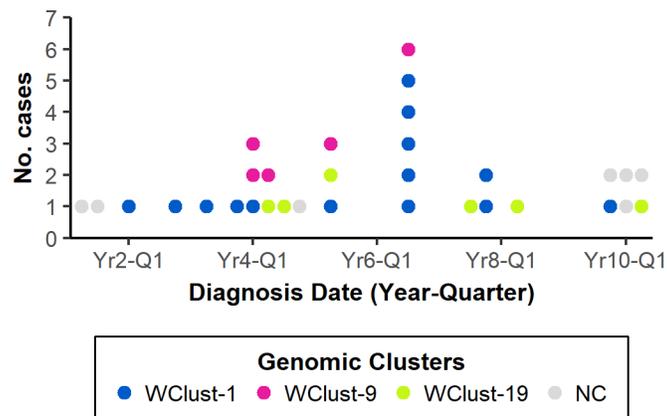


Figure 7-2. Yukon cases over time. Number of tuberculosis cases by year-quarter of diagnosis over a ten-year period in Yukon, Canada. Each circle represents a single case, and colours distinguish the three large clusters identified by a combination of whole genome sequencing and traditional epidemiology. NC (Not Clustered) represents persons with *Mycobacterium tuberculosis* strains unique in Yukon.

7.3.1 Good agreement around clusters and location of TB exposure between methods

Both the CIs+MIRU-VNTR analysis and WGS identified three large clusters (**Figure 7-3**); however, 21/32 (65%) individuals were assigned to one of these three clusters by the team using genotypic data, while WGS placed 25 individuals into these large clusters. Three of the four discordant cases represented scenarios in which the MIRU-VNTR pattern differed from the larger clusters' patterns by a single locus—these were reported as genotypically unique MIRU-VNTR isolates in YT by the laboratory, and led the team to conclude that despite the fact the two of the three discordant cases had epidemiological linkages to known YT cases, these individuals had either acquired TB from an unknown individual in BC ($n = 1$) or another province/territory

($n = 2$). The fourth discordant case had a MIRU-VNTR pattern common to both YT and BC, and while WGS placed this individual with a genomically distinct YT cluster within this MIRU-VNTR group (WClust-1 in **Figure 6-3**), the field team classified this case as having a BC source based on epidemiological information.

Of the remaining seven cases, both teams agreed that two cases were the result of reactivation of LTBI acquired outside Canada; both were persons born outside Canada and involving unique MIRU-VNTR genotypes within YT. For the five other cases, three were genomically clustered with BC isolates (≤ 5 SNVs), one was 26 SNVs from a BC genomic cluster, and one was 18 SNVs from a cluster observed only in YT. The field team classified three of these individuals as having acquired TB from an unknown BC source, thereby agreeing with the genomic assignment. The remaining two were hypothesized by the field team to have acquired their infection within Canada but not YT or BC, which was not supported by the WGS results.

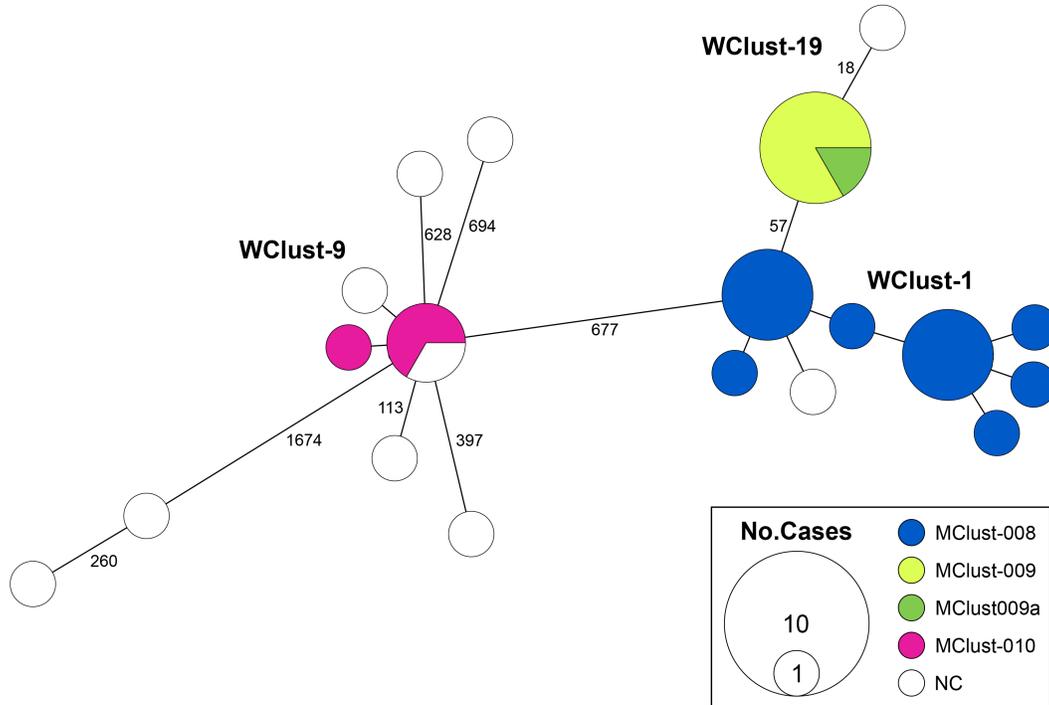


Figure 7-3. Whole genome sequencing-based population structure of Yukon Territory *Mycobacterium tuberculosis* isolates. Minimum-spanning tree based on whole genome sequences of *Mycobacterium tuberculosis* (*Mtb*) isolates from the Yukon Territory (YT), Canada study population ($n = 32$). The size of each circle is proportional to the number of isolates, and circles are coloured to represent the MIRU-VNTR cluster (MClust). Isolates not matching identically at all 24 MIRU-VNTR loci were considered not clustered (NC). Whole genome sequence cluster identifiers (WClustID) are indicated for isolates clustering using a five SNV threshold. The number of SNVs between isolates with >5-SNVs are indicated along the connecting branches.

Ultimately, the two teams agreed on 26/32 (81%) locations of acquisition (**Table 7-1**), with a Cohen's kappa of 0.68 ($p < 0.001$). Concordance was highest amongst individuals belonging to large YT clusters and persons born outside Canada. Qualitative feedback collected at the consensus meeting indicated multiple reasons for conflicting assessments. Unique MIRU-VNTR patterns were cited as a frequent cause—both scenarios in which an isolate's MIRU-VNTR pattern was a single-locus mismatch to an existing cluster, but reported as unique (e.g. WClust-9, **Figure 7-3**) and in which an isolate's MIRU-VNTR pattern was unique to YT but identical to a strain circulating in BC—as was a lack of epidemiological linkages to another YT case.

Examining each method of investigation within the full context of all available data, the genomic-based method had a higher agreement than the field-based approach for identifying connections at a high-level (i.e. provincial/territorial or cluster-level)—30 of 32 (94%, $p = 0.148$) YT cases and all 25 (100%, $p = 0.110$) cases associated with large clusters were correctly categorized (**Table 7-2**). The two discordant genomic-based assignments were the result of low genomic diversity between isolates, and single-locus mismatches to YT MIRU-VNTR patterns leading the team using genomics to assume BC sources.

Table 7-1. Location of TB Acquisition. For each Yukon Territory (YT) *Mycobacterium tuberculosis* isolate ($n = 32$), a pairwise comparison of the two methods used to identify location of tuberculosis acquisition is shown. The four possible categories for location provided to the YT field nurses and BC Centre for Disease Control genomic epidemiologists included YT, British Columbia (BC), Other Province/Territory, and outside Canada.

Genomic Epidemiology	Field Epidemiology				Totals
	YT	BC	Other Prov./Territory	Outside Canada	
YT	17	1	2	0	20
BC	0	7	3	0	10
Other Prov./Territory	0	0	0	0	0
Outside Canada	0	0	0	2	2
Totals	17	8	5	2	32

Table 7-2. High level concordance between methods. Comparison of each method of investigation against the final assignments of tuberculosis source at two levels — province/territory and large cluster.

Method of Investigation	Prov./Territory-Level			Cluster-Level		
	Concordant n (%)	Discordant n (%)	Totals	Concordant n (%)	Discordant n (%)	Totals ^a
Field-based	25 (78)	7 (22)	32	21 (84)	4 (16)	25
Genomic-based	30 (94)	2 (6)	32	25 (100)	0 (0)	25

^aExcluded individuals not linked by both field- and genomic-based methods to a large cluster ($n = 7$).

7.3.2 Low genomic variability within clusters limited of an exact source

Next, the concordance between each team’s identification of a specific source was examined for the subset of cases that acquired TB within YT ($n = 23$). The two teams agreed on a likely source in 13 (57%) instances; of the nine discrepant results, all but one belonged to the largest cluster of cases, WClust-1 (**Table 7-3**). When source case assignments were compared during the in-person consensus meeting, discussion revealed that the team using CIs+MIRU-VNTR data struggled with the complex social network of this cluster, with many connections between individuals, while the team using WGS data were challenged by the minimal genomic diversity between YT isolates (0–4 SNVs). Although the presence of a minority variant in one WClust-1 case (see **Chapter 6**) divided the cluster into two genomically linked sub-clusters, facilitating source identification at the consensus meeting. While there was no strong agreement between the team’s source case assignments, the field-based methods did accurately link individuals to the correct WClust-1 genomic sub-cluster for 11 of 13 persons (85%) with only one individual linked to the incorrect genomic sub-cluster, and a second individual thought to have acquired their infection in BC due to an absence of clear epidemiological connections.

Table 7-3. Concordance between methods at a case-level.
Concordance/discordance between methods of investigation—field- and genomic-based epidemiology—for tuberculosis source case identification, overall and by large cluster.

Characteristic	Concordant <i>n</i> (%)	Discordant <i>n</i> (%)	Totals ^a
Overall	13 (57)	10 (43)	23
Large Cluster			
WClust-1	5 (38)	8 (62)	13
WClust-9	3 (60)	2 (40)	5
WClust-19	5 (100)	0 (0)	5

^aExcluded individuals not assigned a Yukon source by field- and/or genomic-based methods ($n = 12$).

7.3.3 Confidence in correct source identification varied between teams

Next, both teams' level of confidence was examined for each inferred source case/location—not at all, somewhat, very confident or certain. Comparing the confidence category assigned to each inferred source revealed no correlation ($p = 0.365$) between the two teams. The team using genotyping and contact data on average reported higher levels of certainty in their source ascertainment ($p = 0.007$) (**Figure 7-4**). Differences were also noted within and across the large clusters (**Figure 7-5**). The largest genomic cluster (WClust-1) had the widest distribution of confidence in source case ascertainment; conversely, participants reported higher confidence in inferred sources in the smaller clusters (WClust-9; WClust-19), particularly the team using CIs+MIRU-VNTR data, who reported "very confident" or "certain" for all source cases identified.

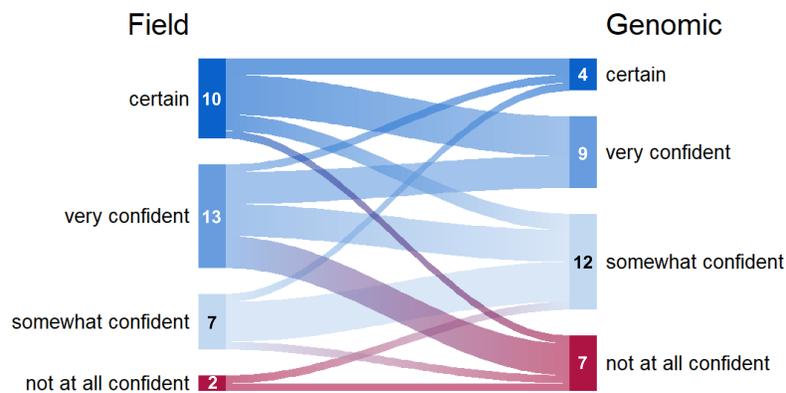


Figure 7-4. Certainty assigned to identified sources. Relationship between degree of certainty assigned to each source case/location identified by field- and genomic-based methods. Link widths are proportional to the number of cases, which are indicated in margins.

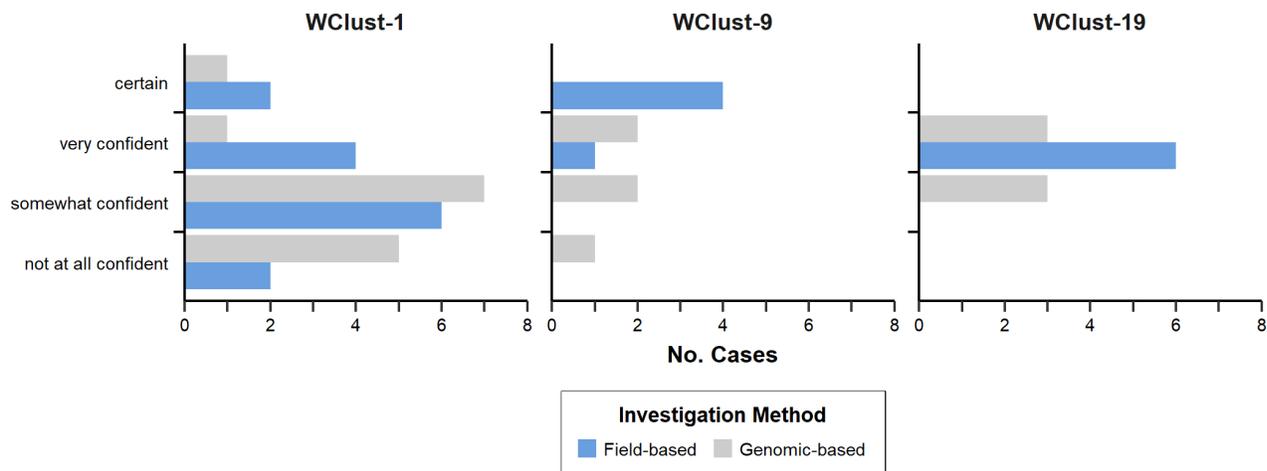


Figure 7-5. Frequency of certainty categories assigned for each source identified, divided by cluster.

During the in-person consensus meeting, the genomic, clinical, and contact investigation data were combined and discussed at length to determine the most plausible source for each individual. Overall, YT-resident sources were assigned to 23 individuals, and the degree of certainty assigned to each source (field-based: $n = 20$; genomic-based: $n = 22$) identified during the independent investigations was examined. It was found that where there was a high level of confidence (i.e. "very confident" or "certain") the correct source had been identified in 100% of cases using either investigation method (field-based: $n = 12$; genomic-based: $n = 5$), **Table 7-4**. Where there was a lower level of certainty for the identified source, only five of eight (62%, field-based), and 10 of 17 (59%, genomic-based) correctly identified the source ($p > 0.05$).

Table 7-4. Accuracy of source case identification. Frequency of "correct" tuberculosis source case identification by levels of confidence for each investigation method.

Level of Confidence	Field-based ^a n (%)		Genomic-based ^b n (%)	
	Correct Source	Incorrect Source	Correct Source	Incorrect Source
Very Confident/Certain	12 (100)	0 (0)	5 (100)	0 (0)
Somewhat/Not at All Confident	5 (62)	3 (38)	10 (59)	7 (41)

^aExcluded individuals not assigned a Yukon source by field-based methods ($n = 12$).

^bExcluded individuals not assigned a Yukon source by both field- and genomic-based methods ($n = 10$).

7.3.4 Preference for genomics over genotyping

Each member of the YCDC TB program team ($n = 4$) completed an online survey at the outset of the study in which they were asked about their role in TB prevention and care and their knowledge of genotyping methods and use in TB investigations. All respondents were engaged in direct care and treatment of persons with TB, including the collection of personal information, supervising daily medication doses, CIs, and program oversight, and three of four spend an average of >60% of their week on TB-related activities. All respondents have a background in nursing with approximately 13 to 35 years of experience—most of which in rural and remote communities.

Three of the four team members had heard of MIRU-VNTR prior to this study, through presentations, conferences, and/or journal articles. Only one respondent reported using MIRU-VNTR information in their daily work. None indicated that they had received formal training in the use and interpretation of MIRU-VNTR in TB investigations, although three of four were aware that MIRU-VNTR data were available for their cases. No respondents reported complete confidence in using MIRU-VNTR data for their investigations, and none had used MIRU-VNTR to inform their TB investigations prior to this study.

At the conclusion of the in-person consensus meeting, a semi-structured group interview was conducted to collect qualitative feedback regarding the use of molecular/genomic data in this setting. Two main themes emerged as detailed on the next page.

Key themes from a semi-structured group interview with the field team regarding the use of tuberculosis molecular/genomic data.

1. **The accuracy of genomics over genotyping.** The team reported that MIRU-VNTR genotyping data conflicted with known epidemiological connections in a number of instances, whereas genomics identified clusters more closely aligned with the epidemiological data, and provided some novel insight.

“I’m liking MIRU a little less”

“The MIRU can be helpful or not helpful”

“To have had the WGS data, would have saved many hours of discussion—would have helped to focus the discussion by narrowing the list of potential sources”

2. **Program assessment.** Participants acknowledged that genomic epidemiology provided new insights into transmission patterns and saw WGS as a way to assess the effectiveness of treatment and prevention programs, including screening and prophylaxis.

“Many of these confirmed our suspicions”

“It was nice to know this was a reactivation and not a contact of a missed source”

“Small case load means few people working on TB, and we need to focus limited resources on the highest risk contacts.”

“...prophylaxis could have prevented the cluster”

Overall, participants viewed genomic epidemiology as a useful tool to streamline investigations, particularly in differentiating LTBI reactivation from recent transmission, but not essential to their current practices, instead noting it would most be useful for program assessment. The team found that WGS results were useful for confirming probable source cases and ruling out local transmission. MIRU-VNTR data was cited as a source of frustration where it did not align with the epidemiology. Improved communication around how to interpret closely related MIRU-VNTR patterns, as well as the limitations of genotyping, was strongly recommended. In a post-meeting follow-up online survey (**Appendix Table A-3**), respondents reiterated the themes from the in-person group interview by highlighting their preference for WGS over MIRU-VNTR, with qualitative feedback such as *“WGS provides a clearer picture than MIRU-VNTR of what is happening in terms of transmission.”* Additionally, respondents noted that WGS highlighted some gaps in knowledge or what may have been missed during contact tracing, supporting the idea of using WGS towards program assessment. When asked if they felt more confident using WGS data following this study, all stated that they were considerably more confident and would like to have genomic data for all cases. The team also indicated they would be open to further training in the interpretation of genomic data, with in-person training preferred over an instruction manual or instructional videos.

7.4 Discussion

In this study, both the added value of using genomic epidemiology in settings with rich field epidemiological data, as well as the knowledge, attitudes and practices around the use of molecular and genomic data for TB case investigations was examined. Comparing the traditional approach of inferring TB transmission from genotyping and contact investigation data to the use of genomics with limited case-level data revealed that WGS more accurately identified connections between cases at a high level, such as cluster membership, but that the data from CIs was integral to identifying source cases at an individual-level, particularly within large clusters.

In certain settings, genotyping by 24-locus MIRU-VNTR has been reported as having high discriminatory power and good concordance with known epidemiological linkages.^{118,306,393} However, technical issues with particular loci, rendering them untypable, can make cluster assignment challenging—these patterns are often assigned their own unique identifier, and obscure the potential linkage between isolates.^{76,108,394} *Mtb* isolates with single-locus mismatches have been shown to be linked by both epidemiology and genomics.^{112,149,306,354,355} In this study, these “falsely unique” MIRU-VNTR patterns complicated the interpretation of contact investigation data, with the true nature of clustering only revealed through the higher-resolution genomic approach. Given that WGS may not be available to all TB programs, it is recommended that laboratories reporting MIRU-VNTR data include information not just on identical patterns, but also closely related patterns that might suggest a larger cluster. The survey results also indicated that there is a substantial gap in training end-users to interpret genotyping data, suggesting that laboratories might consider including some interpretive commentary on their genotyping reports beyond simply a pattern and a cluster identifier.

As expected, genomic data coupled to basic clinical and epidemiological data was able to identify clusters and infer some potential sources, but it was only when it was combined with extensive contact investigation data that a more comprehensive picture of TB transmission began to emerge. This highlights the importance of engaging both the laboratory and public health nursing and epidemiology staff in the joint interpretation of genomic epidemiology data. In remote northern settings with extensive person-to-person transmission,^{166,379} the minimal

genomic variation observed means that data from CIs is integral to understanding local epidemiology, and that enhanced investigation questionnaires, as recently used in a United Kingdom study, can establish epidemiological connections between individuals that would have otherwise not been linked.¹¹⁸ During discussions with the YCDC public health team, they noted that had WGS data been available during CIs, more focused questioning likely would have uncovered some missed connections and would have helped to confirm/refute tenuous linkages, saving time and resources. Discussions also revealed a strong preference for WGS over MIRU-VNTR to support CIs, and identified program assessment as an important secondary use for WGS data.

This study also alluded to the importance of sharing molecular epidemiology data across jurisdictional boundaries. Genotyping results are routinely reported at the provincial/territory level, but information on the presence of a pattern in another jurisdiction may not always be provided. Here, the YCDC team did not have access to molecular data from BC cases prior to this study, and during routine CIs were unaware that several cases with MIRU-VNTR patterns unique to YT were actually members of genotypic clusters comprising multiple BC cases.

A major strength of the present study was the availability of a small, well-characterized population, particularly with the long service of several of the nurses involved in the study who have considerable experience within the community. A limitation of the comparison between investigation methods was that the team using data from CIs could make connections between culture-positive and -negative cases; however, the molecular and genomic analyses were limited to culture-positive cases and may have resulted in missed linkages between individuals.

This study highlights the need to better integrate laboratory, clinical, and epidemiological data to more comprehensively describe TB epidemiology in a given setting, including using higher-resolution genomic approaches where possible, providing better interpretation of MIRU-VNTR data when WGS is not available, and bringing individuals together for collaborative discussion of cases and clusters. Through a molecular-informed, enhanced contact investigation approach, it is believed that TB programs might better focus their resources and avoid missed opportunities

for intervention, thereby limiting new transmissions. For this to occur, communication is key. Given the dynamic and complex nature of genomic and contact investigation data, regular review of cases through in-person meetings and training in interpretation is recommended. Genomics also has the potential to aid in TB program evaluation, and as the technique becomes more commonplace, TB laboratories and prevention and care programs must work together to jointly assess the impact of this emerging epidemiological approach.

Chapter 8: Whole Genome Sequencing as a Tool to Understand and Quantify Active Tuberculosis Arising from Local Transmission

8.1 Background

While TB in low-incidence settings such as Canada is thought to largely result from reactivation of latent TB infection (LTBI) in migrants to Canada—67% of diagnoses occur in non-Canadian-born (nCB) persons⁶—quantifying the actual contribution of this LTBI reactivation versus locally acquired infection has been difficult. Genotyping methods, such as 24-locus mycobacterial interspersed repetitive unit–variable-number tandem repeat (MIRU-VNTR), have provided some insight by identifying unique strains presumed to represent LTBI reactivation.^{101,342,395} However, the increasingly frequent application of whole genome sequencing (WGS) to *Mycobacterium tuberculosis* (*Mtb*) isolates has called into question genotyping’s utility for identifying local transmission,^{336,354,396–398} particularly amongst non-Euro-American lineage *Mtb* strains, where the method over-estimates clustering.³¹⁸ Beyond improved resolution for defining clusters, genomics also permits the inference of both the timing and direction of person-to-person spread.^{74,144,149,158,174}

British Columbia (BC) has committed to reducing TB incidence rates by 50% by the year 2022 as part of its Provincial TB Strategy.¹⁴ Reducing local transmission through targeted, evidence-based interventions has been identified as an important objective within the strategy. This necessitates first determining the proportion of active TB disease due to local, i.e. BC-based, transmission. Then filling knowledge gaps in our understanding of this transmission, particularly with respect to describing general characteristics of transmission networks, identifying the drivers of large outbreaks, as well as enumerating the clinical, demographic, and epidemiological risk factors associated with persons who give rise to multiple secondary cases (transmitters or super-spreaders) versus non-transmitters. Here, the genomic epidemiology of TB transmission in BC over a ten-year period is described, as a first step in developing bespoke interventions aimed at reducing local transmission within the province.

8.2 Methods

8.2.1 Study population

Mtb specimens and referred-in cultures for the province are received by the British Columbia Centre for Disease Control (BCCDC)'s Public Health Laboratory (BCPHL), which oversees routine diagnostics, phenotypic drug sensitivity testing, and 24-locus MIRU-VNTR genotyping. TB prevention programs, routine surveillance, and patient care are led by the BCCDC's Provincial TB Services program. The study population included all persons with culture-confirmed TB residing in BC whose first *Mtb* isolate was received by the BCPHL from 2005 through 2014, representing 79.5% of TB cases reported to TB Services during this period.

8.2.2 Case data

Individual-level clinical and demographic data variables were extracted from BCCDC's electronic medical registry, Integrated Public Health Information System (iPHIS). To support transmission analyses, further case information was obtained as needed through chart review of physician narratives and contact investigation records in iPHIS. For demographic analyses, community type was determined using the population density of the geographic service area in which each case resided—metro (>190,000), urban/rural (40,001–190,000), rural (10,001–40,000), and remote (\leq 10,000). Census dissemination areas (DA), based on postal codes for each case, were linked to the 2006 Canadian Marginalization Index (CAN-Marg)³¹⁶ to determine the deprivation index quintile representing the relative socioeconomic disadvantage of a DA compared to the rest of Canada, (quintile 1: least deprived, quintile 5: most deprived). To protect privacy, some data is suppressed at the DA level by Statistics Canada, resulting in a small number of unlinked records for individual deprivation indices.

8.2.3 Laboratory analysis

2,303 *Mtb* isolates were revived from BCPHL's archival stocks, DNA was extracted, and genotyped using 24-locus MIRU-VNTR genotyping as previously described (**Chapter 3**). Phenotypic drug susceptibility testing (DST) results for first-line antibiotics—isoniazid (INH), rifampin (RIF), ethambutol (ETB), and streptomycin (SM)—with additional data for

pyrazinamide (PZA) in multi-drug resistant isolates, were available for each isolate through routine testing on the BACTEC MGIT 460 or 960 (Becton-Dickinson, Sparks, MD) interpreted according to Clinical and Laboratory Standards Institute (CLSI) recommendations.³⁹⁹ Multi-drug-resistant (MDR) tuberculosis was defined as an isolate resistant to INH and RIF.

Assuming that isolates with a unique MIRU-VNTR genotype would also be unrelated to another BC isolate by WGS, the subset of genotypically clustered isolates representing possible local transmission were sequenced. The 974 MIRU-VNTR-clustered isolates were sequenced using 125 bp paired-end reads on the Illumina HiSeq 2500 platform at the Michael Smith Genome Sciences Centre (Vancouver, BC). An additional 247 *Mtb* isolates of special interest were also sequenced. These genotypically unique isolates were selected by meeting any or all of the following criteria: (i) isolates from Canadian-born persons ($n = 105$), (ii) isolates matching at 23 of 24 MIRU-VNTR loci to a genotypic cluster likely to represent local transmission ($n = 37$), (iii) isolates from a culture-positive TB reoccurrence either during the 2005–2014 study period or beginning within the five years prior to the study period ($n = 51$), and (iv) isolates phenotypically resistant to one or more first-line antibiotic ($n = 118$).

8.2.4 WGS analysis and genomic clustering

The resulting FASTQ files were analyzed using a pipeline developed by Oxford University and Public Health England.²⁶⁷ Reads were aligned to the *Mtb* H37Rv reference genome (GenBank ID: NC000962.2), and after masking for low complexity regions an average of 92% of the reference genome was covered. Single nucleotide variants (SNVs) were identified across all mapped non-repetitive sites and concatenated SNVs were used to construct a maximum-likelihood phylogenetic tree in RAxML 8.2.10³⁷¹, using the GTRGAMMA model and 200 bootstrap replicates. Major lineage was predicted for each isolate, first by using TB-Insight's³²⁹ CBN method with MIRU-VNTR data as input, and second by lineage-specific SNVs³⁷² for those isolates that were sequenced. FASTQ files for all genomes are available at NCBI under BioProject PRJNA413593 and PRJNA49659.

While several studies^{73,149,158,165,166} have used SNV thresholds—typically 5–12 SNVs—as a bound for recent transmission, thresholds are sensitive to SNV filtering approaches, and a hard

cut-off of five or 12 SNVs might exclude some cases of interest, for example, reactivation cases of a strain endemic to BC. Thus, it was decided to initially define locally acquired infections as those in genomic clusters within 0–20 SNVs of another study isolate, with further refinement of the threshold based on clinical and demographic analysis. In a sensitivity analysis, clustering estimates were recalculated using thresholds of both 5 and 12 SNVs.

8.2.5 Transmission across population groups

All genomic clusters containing both CB and nCB persons were analyzed to identify possible transmission between individuals in these population groups. A combination of genomic and epidemiological information obtained from case-record review was used to determine the most likely source of an individual's infection. A diagram was constructed to depict the number and direction of transmission events between CB and nCB persons. The strength of the epidemiological linkage is indicated as: (1) household contacts, (2) known—one or both individuals named the other and do not reside in the same household, (3) probable—individuals did not name one another yet share epidemiological characteristics along with overlapping geographical locations during the likely infectious period of the source, and (4) unknown—individuals do not have clear connections to one another.

8.2.6 Tuberculosis reoccurrences

Mtb isolates from individuals diagnosed with TB from 2005 through 2014, and who either had a second episode of culture-confirmed TB within the study period or had a previous TB episode of culture-confirmed TB in 2000–2004, were analyzed to differentiate between relapse (second episode >6 months after last treatment) and exogenous reinfection. SNV profiles for each isolate were examined within the context of the larger study population, in addition to case-level characteristics of each individual obtained through case record review, including first episode treatment status, travel to endemic country of origin between episodes, and health-related risk factors such as HIV and diabetes. Descriptive statistics were calculated for cases determined to be either relapse or reinfection.

8.2.7 Characterization of large clusters and transmission reconstruction

To characterize ongoing endemic transmission representing the greatest potential for public health intervention, the epidemiological characteristics of genomic clusters (20-SNV threshold) with ≥ 10 isolates were described. Additionally, transmission pathways were reconstructed for one of the largest clusters not previously described in the literature. To do this SNV profiles were analyzed within the context of case-level clinical, demographic and contact investigation information gained through detailed case-record review. An initial network was constructed using concatenated SNVs which was then refined with epidemiological data to represent the most likely transmission pathway. Where transmission likely occurred prior to the study period a node was placed in the network to indicate such an event. Epidemiological linkages were classified by the strength of evidence connecting cases. A linkage was categorized as “known” if either the source or contact named the other. “Probable” linkages represented individuals that spent time in the same locale (e.g. shelter) which overlapped the likely infectious period of the source and shared unique characteristics within the cluster (e.g. same nCB country of birth within a CB cluster) or social/behavioural characteristics (e.g. substance use), and where neither identified the other as a contact. Where none of these criteria applied the epidemiological association between genomically linked isolates was classified as “unknown”.

8.2.8 Statistical analysis

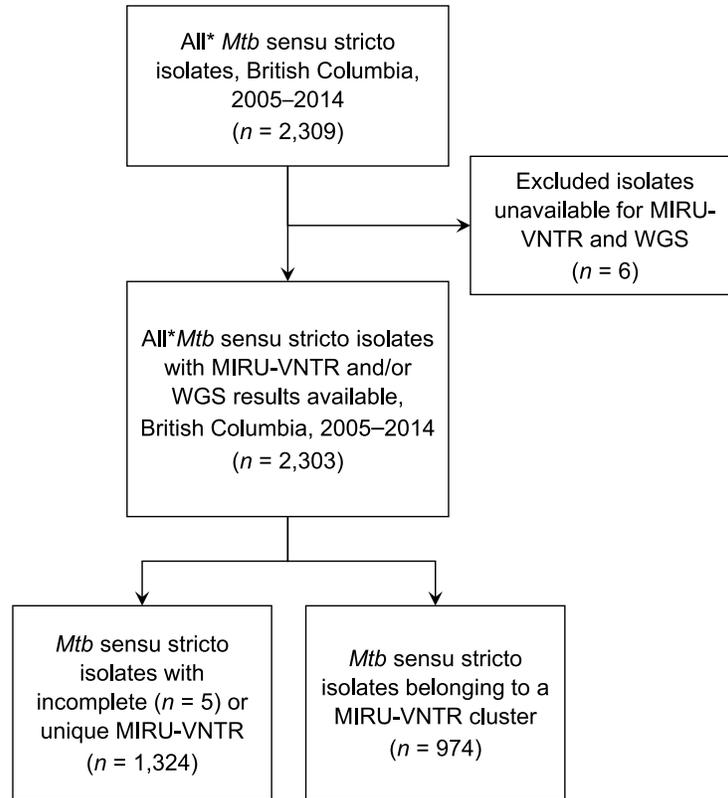
A genotypic cluster was defined as ≥ 2 individuals with identical MIRU-VNTR results and a unique identifier (MClustID) was assigned to each cluster. A unique WGS cluster identifier (WClustID) was given to isolates which were within 20 SNVs of another individual’s isolate within the study. Regression models were constructed under the assumption that isolates that were not selected for sequencing would be >20 SNVs from other study isolates and therefore were included as genomically unique cases. A multivariable logistic regression model was constructed to estimate the odds ratio (OR), adjusted odds ratio (aOR) and 95% confidence interval (CI) for the distribution of cases by cluster status (clustered/non-clustered) according to birthplace and other clinical and demographic variables using backward elimination of factors identified in univariable analysis ($p < 0.20$), and Akaike’s Information Criterion (AIC) minimisation.³³¹ A similar model was constructed to examine the characteristics where three

different SNV thresholds for clustering were used (5, 12 and 20 SNVs). Given that the substance use variables had >5% missing values, Little's test³³² was used to determine whether data was missing completely at random (MCAR). A significant result ($p < 0.001$) indicated the missing values were not MCAR, and a complete-case analysis was used for the logistic regression models. Cases excluded from the analysis due missing data were compared to those remaining in the analytic sample using Chi-square tests to assess potential bias due to missing data.

Descriptive statistics were calculated to characterize all large (≥ 10 isolates) genomic clusters according to birthplace, age, gender, substance use, housing status, community type and number of cases with known epidemiological linkages to others in the cluster. Under-housed was defined as an individual with no fixed address or someone living in a homeless shelter, group home, or residing in single-room occupancy (SRO) housing. All statistical analyses were executed in R (v3.4.1).

8.3 Results

From 2005 through 2014, 2,309 *Mtb* isolates were received at BCPHL for routine testing. These represent the first isolate of each culture-positive TB case (excluding subsequent isolates from a relapse) and 79.2% of all TB diagnoses in BC during this time period. Six isolates were unavailable and excluded from the study, leaving a total study sample of 2,303 *Mtb* isolates (**Figure 8-1**). The characteristics of the study population are described in **Table 8-1**. Briefly, 30.7% of the study population were aged 35–54, with higher proportions of males (58.2%), individuals residing in a metro area (76.5%), and individuals born outside Canada (73.6%). The majority of nCB individuals had immigrated ≥ 5 years prior to diagnosis (71.5%), and 87.5% of nCB were born in Asia (87.5%). Clinically, 84.1% of the study cohort had respiratory disease, and 62.1% were smear-positive. Chest radiographs identified cavitory TB in 13.9% of individuals. HIV-positivity was reported for 4.5% of the study population, and the proportion of substance use was recorded as 8.8% for illicit drug use and 9.6% alcohol misuse. With respect to marginalization, 43.9% of individuals resided in the most deprived quintiles (≥ 4 on the material deprivation index).



**Only the first isolate was included for cases with subsequent isolates representing a TB relapse during the study period*

Figure 8-1. Study sample inclusion/exclusion criteria. Number of *Mycobacterium tuberculosis* (*Mtb*) isolates at each stage of the study sample selection.

Abbreviations: WGS, whole genome sequencing; MIRU-VNTR, Mycobacterial Interspersed Repetitive Units–Variable Number Tandem Repeats.

Table 8-1. Characteristics of the study sample.Demographic and clinical characteristics of culture positive TB cases, British Columbia 2005–2014 (*n* = 2,303)^a.

Characteristic	No. Cases (%)
Age, years	
0–14	32 (1.4)
15–34	504 (21.9)
35–54	708 (30.7)
55–74	588 (25.5)
75+	471 (20.5)
Gender ^b	
Male	1339 (58.2)
Community type	
Metro	1762 (76.5)
Urban/Rural	334 (14.5)
Rural	174 (7.6)
Remote	33 (1.4)
Birthplace ^c	
Canada	592 (26.4)
Non-Canadian-born continent ^d	
Asia	1444 (87.5)
Africa	80 (4.8)
Europe	69 (4.2)
Americas	46 (2.8)
Oceania	11 (0.7)
Time in Canada ^e	
< 5 years	458 (28.5)
≥ 5 years	1148 (71.5)
Disease Site	
Respiratory	1776 (77.1)
Non-Respiratory	366 (15.9)
Respiratory + Non-Respiratory	161 (7.0)
Respiratory ^f smear	
Positive	1159 (62.1)
Cavitary disease	
Yes	319 (13.9)
Drug susceptibility	
MDR	19 (0.8)
INH-R (non-MDR)	174 (7.6)

Table 8-1 *Continued from previous page*

Characteristic	No. Cases (%)
HIV	
Positive	104 (4.5)
Negative	1810 (78.6)
Unknown	389 (16.9)
Illicit drug use	
Yes	203 (8.8)
No	1616 (70.2)
Unknown	484 (21.0)
Alcohol misuse	
Yes	220 (9.6)
No	1622 (70.4)
Unknown	461 (20.0)
Material deprivation ^g	
Quintile 1 (least)	275 (12.6)
Quintile 2	420 (19.2)
Quintile 3	532 (24.3)
Quintile 4	530 (24.2)
Quintile 5 (most)	429 (19.7)

Abbreviations: HIV, human immunodeficiency virus; INH-R, isoniazid resistant; MDR, multi-drug resistant tuberculosis (resistant to isoniazid and rifampin).

^aPercentages have been rounded and may not total 100%.

^bOne transgender/gender-unknown individual excluded from analysis.

^cData unavailable in 57 cases.

^dData unavailable in 4 cases.

^eData unavailable in 48 cases.

^f“Other respiratory” sites (e.g. pleura) were excluded.

^gData unavailable in 116 cases.

8.3.1 Whole genome sequencing reduces the local transmission estimate

The retrospective 24-locus MIRU-VNTR genotyping of 2,303 *Mtb* isolates representing >99% of all culture-positive TB cases diagnosed in BC between 2005–2014 was described in **Chapter 3**, in which 42.4% of isolates were clustered and potentially representative of recent transmission. By WGS (**Figure 8-2**), only 594 (25.8%) were clustered at a 20-SNV threshold, substantially lowering the estimate of local transmission within BC. The reduction of the clustered proportion was largely due to nCB isolates—only 24.2% of MIRU-VNTR clustered isolates from nCB persons were genomically clustered, in contrast to 94.5% of isolates from CB persons. Study isolates belonged to 93 distinct genomic clusters, ranging in size from 2–72 isolates (**Figure 8-3**). While 57.0% of clusters consisted of just two persons, individuals in large clusters (≥ 10 cases/cluster) accounted for 61.4% (365/594) of all clustered cases. Using the “ $n - 1$ ” method,⁸⁸ in which the first isolate in each genomic cluster is not counted, the proportion of BC’s TB cases resulting from local transmission was estimated at 21.8%.

The clustered proportion remained stable when the SNV thresholds for clustering were reduced from 20 SNVs to 12 and 5 SNVs. Thirty-three cases were no longer considered clustered at the 12-SNV threshold, reducing the clustered proportion from 25.8% to 24.4% (561/2,303), while further reduction of the threshold to 5 SNVs left 511 (22.2%) isolates clustered.

Plotting the minimum pairwise SNV distance for each sequenced isolate within the study period—the majority of which were genotypically clustered—revealed that 74.4% (406/546) of isolates obtained from Canadian-born persons fell within 0–5 SNVs of another study isolate. In contrast, only 15.6% ($n = 96/617$) of isolates from nCB persons were similarly clustered. The majority (68.7%) of nCB persons had isolates >50 SNVs from the nearest isolate in BC (**Figure 8-4**).

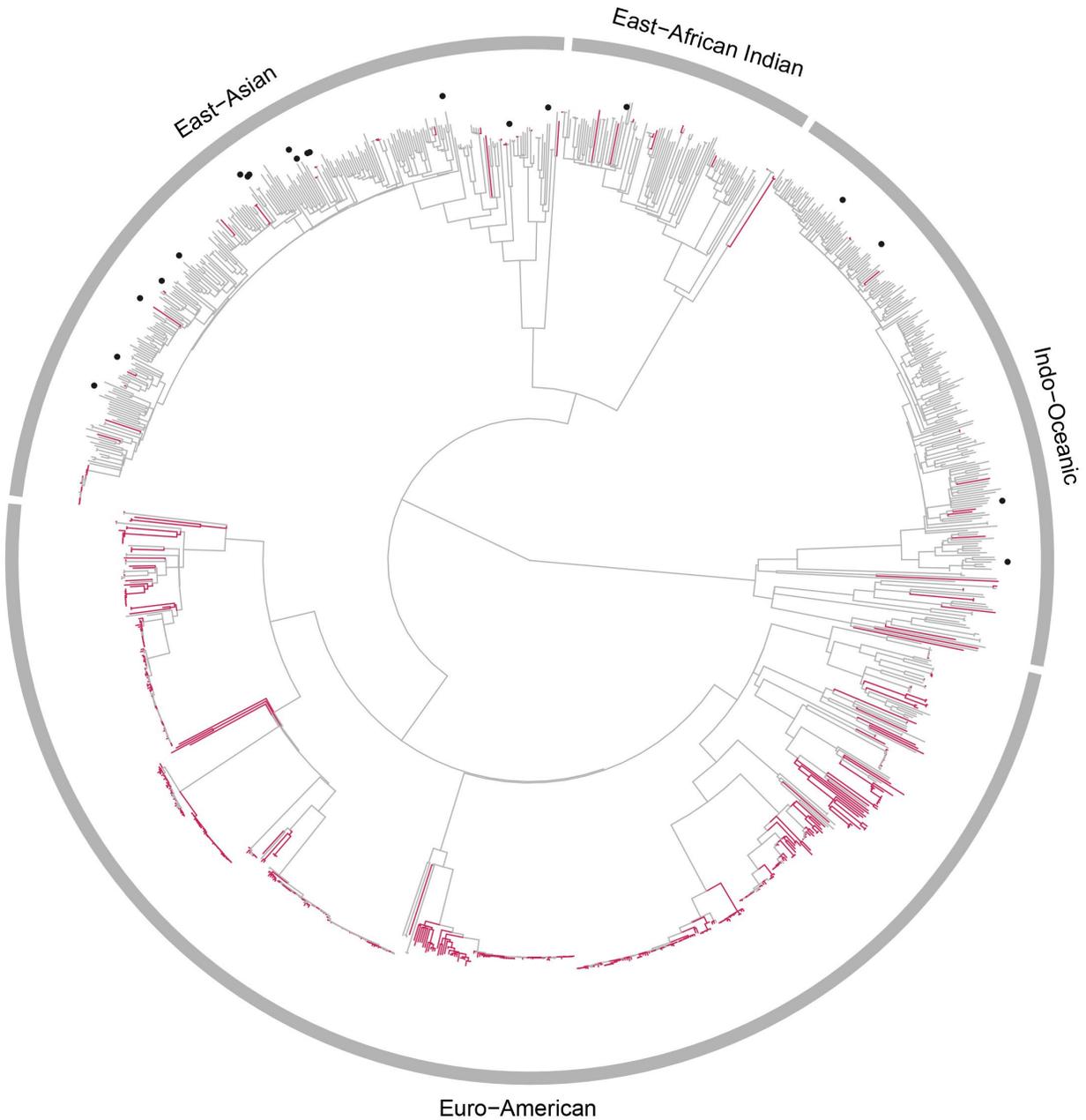


Figure 8-2. Maximum-likelihood phylogenetic tree of study isolates. The tree was constructed from concatenated single nucleotide variant alignments derived from whole genome sequencing of *Mycobacterium tuberculosis* isolates ($n = 1,221$). Isolates from Canadian-born persons are indicated in red (—). Multi-drug resistant isolates—those resistant to isoniazid and rifampin—are indicated by a circle (●). The outer ring indicates the lineage.

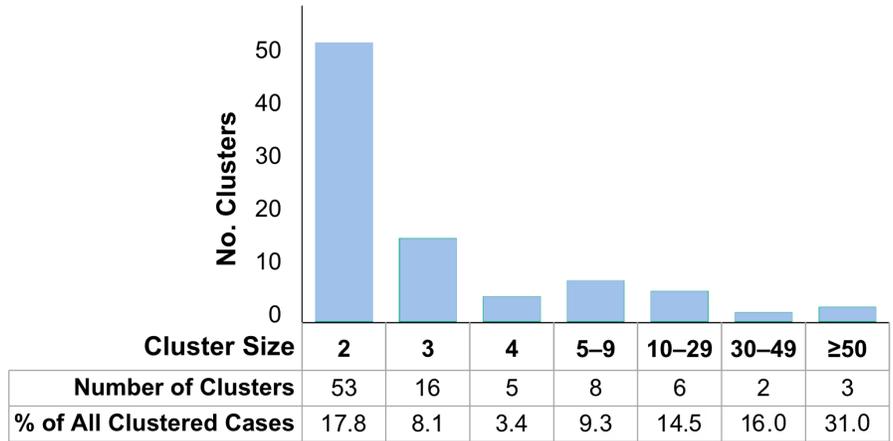


Figure 8-3. Genomic cluster sizes. Number of genomic clusters by size, and the proportion of all clustered isolates represented at each cluster size. A threshold of 20 single nucleotide variants was used to define a cluster ($n = 93$), British Columbia, 2005–2014.

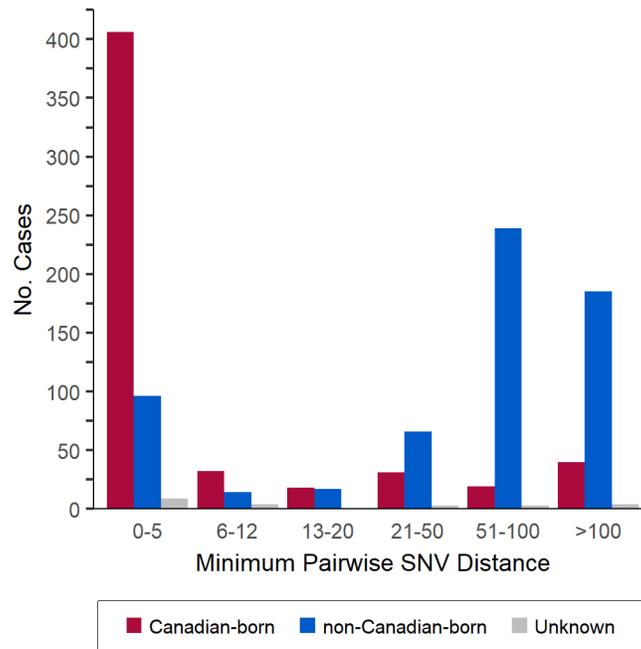


Figure 8-4. Pairwise SNV distances between study isolates. Minimum pairwise single nucleotide variant (SNV) distance between isolates, coloured according to birthplace—red for Canadian-born, blue for non-Canadian-born, and grey for persons with an unknown birthplace. Of the 1,221 isolates that were sequenced, 24 isolates received prior to 2005, and 11 isolates representing a TB relapse were excluded from this figure.

8.3.2 Transmitted *Mtb* isolates belong largely to the Euro-American lineage

Clustering proportions varied significantly across lineages—54.0% of Euro-American lineage strains were genomically clustered, in contrast to 12.6% of East-Asian, 4.5% of Indo-Oceanic and 7.9% of East-African Indian lineage isolates ($p < 0.001$). Restricting the analysis to consider only genomically clustered isolates, the Euro-American lineage dominated with 79.6% of clustered isolates belonging to this lineage ($p < 0.001$). All but one of BC's 11 large genomic clusters (11–72 isolates/cluster) belonged to the Euro-American lineage and were comprised of predominantly CB persons. The single non-Euro-American cluster, WClust-6, belongs to the East-Asian lineage and the majority of cases—88.9% (16/18)—were Canadian-born persons, with nine (50%) known to have connections to a highly marginalized urban area.

A maximum-likelihood tree (**Figure 8-2**) reveals the population structure of BC's *Mtb* isolates. Isolates from CB persons were concentrated in the Euro-American *Mtb* lineage (480 of 546 isolates, 87.9%). Of the 85 CB persons with a non-Euro-American isolate and excluding those belonging to the known transmission cluster WClust-6, it was determined that the *Mtb* lineage aligned with the travel history, ethnic community, or known contact to a nCB case for 48/69 (69.6%) individuals. For the remaining 21 (30.4%) individuals, there was no documented connection to the geographic region or ethnic community expected according to the *Mtb* lineage of their isolate. Of note, 61.9% (13/21) of these cases with an unexplained non-Euro-American lineage isolate resided in the Greater Vancouver Region—an ethnically diverse metropolitan area of BC.

8.3.3 Risk factors for local transmission

Due to missing values for certain risk factor variables, 537 records were excluded from logistic regression analysis, resulting in an analytic sample of 1,766 cases. Missing data was more often associated with particular regions of BC, with rural Health Service Delivery Areas (HSDAs) having significantly higher proportions of missing data for substance use variables (illicit drug use: 35.4% [96/271], alcohol misuse: 30.3% [82/271]), as compared to urban-based HSDAs, for which 19.1% (388/2,032) of illicit drug use and 18.7% (379/2,032) of alcohol misuse fields had

missing values ($p < 0.001$). Because data missingness was not likely associated with the variables themselves or other case characteristics and was instead related to data entry, the decision was not to impute missing values and instead use a complete-case analysis for logistic regression. Given that the demographics of individuals in rural HSDAs differ from those in urban regions, cases excluded from the analytical sample were more likely to be older ($p < 0.001$), male ($p = 0.034$), CB ($p < 0.001$), and have exclusively respiratory TB ($p = 0.005$).

Using a 20-SNV threshold, the proportion of culture-confirmed TB arising from local transmission was calculated as 76.7% (454/592) for CB persons and 7.7% (127/1654) for nCB individuals, and with the complete-case logistic regression analysis CB persons were found to have 19.7 (95%CI: 13.9–28.0) times the odds of belonging to a genomic cluster as compared to non-Canadian-born persons, after adjustment for age, gender, anatomical disease site, and substance use (**Table 8-2**). Other significant risk factors for clustering indicative of local transmission included respiratory disease (aOR 2.4, 95%CI: 1.5–4.0), illicit drug use (aOR 3.8, 95%CI: 2.0–7.2) and alcohol misuse (aOR 5.0, 95%CI: 2.8–8.9). In a sensitivity analysis examining factors associated with clustering, when the multivariable regression model was fitted with lower SNV clustering thresholds (12 and five SNVs), estimated odds ratios did not change appreciably (**Table 8-3**).

Table 8-2. Genomic clustering logistic regression. Distribution and multivariable analysis of factors associated with genomically clustered *Mycobacterium tuberculosis* isolates (≤ 20 SNVs) and unique isolates (>20 SNVs), British Columbia 2005–2014 ($n = 1,766$).

Characteristic	Clustered <i>n</i> (%)	Unique <i>n</i> (%)	Clustered vs. Unique OR (95%CI)	Clustered vs. Unique aOR ^a (95%CI)
Age, years				
0–14	11 (37.9)	18 (62.1)	1.7 (0.8–3.8)	1.0 (0.4–2.8)
15–34	104 (26.0)	296 (74.0)	Reference	Reference
35–54	207 (36.3)	364 (63.7)	1.6 (1.2–2.1)	1.2 (0.8–1.8)
55–74	93 (19.8)	377 (80.2)	0.7 (0.5–1.0)	0.5 (0.3–0.8)
75+	14 (4.7)	282 (95.3)	0.1 (0.1–0.3)	0.2 (0.1–0.3)
Gender				
Male	278 (27.6)	729 (72.4)	1.5 (1.2–1.9)	1.1 (0.8–1.5)
Female	151 (19.9)	608 (80.1)	Reference	Reference
Birthplace				
Canada	321 (77.5)	93 (22.5)	39.8 (29.4–53.8)	19.7 (13.9–28.0)
Outside Canada	108 (8.0)	1244 (92.0)	Reference	Reference
Disease Site				
Resp.	370 (27.5)	976 (72.5)	3.0 (2.1–4.4)	2.4 (1.5–4.0)
Non-Resp.	34 (11.2)	269 (88.8)	Reference	Reference
Resp. + Non-Resp.	25 (21.4)	92 (78.6)	2.1 (1.2–3.8)	1.8 (0.8–3.9)
Illicit drug use	157 (90.8)	16 (9.2)	47.7 (28.0–81.0)	3.8 (2.0–7.2)
Alcohol misuse	154 (84.6)	28 (15.4)	26.2 (17.1–40.0)	5.0 (2.8–8.9)

Abbreviations: SNVs, single nucleotide variants; OR, odds ratio; CI, confidence interval; aOR, adjusted odds ratio; Resp., respiratory; Non-Resp., non-respiratory.

^aAdjusted for age, gender, birthplace, disease site, illicit drug use and alcohol misuse.

Table 8-3. Logistic regression for risk factors for genomic clustering— various thresholds. Multivariable analysis of factors associated with clustered *Mycobacterium tuberculosis* isolates using three different upper thresholds for clustering, 20 SNVs, 12 SNVs, and 5 SNVs versus unique isolates (>20 SNVs or unique MIRU-VNTR for isolates not sequenced), British Columbia 2005–2014 ($n = 1,766$).

Characteristic	Clustered vs. Unique aOR (95%CI) ^a		
	20 SNV	12 SNV	5 SNV
Age, years			
0–14	1.0 (0.4–2.8)	1.2 (0.4–3.3)	1.6 (0.6–4.5)
15–34	Reference	Reference	Reference
35–54	1.2 (0.8–1.8)	1.2 (0.8–1.9)	1.2 (0.8–1.9)
55–74	0.5 (0.3–0.8)	0.5 (0.3–0.9)	0.5 (0.3–0.9)
75+	0.2 (0.1–0.3)	0.1 (0.1–0.3)	0.1 (0.1–0.3)
Gender			
Male	1.1 (0.8–1.5)	1.2 (0.8–1.7)	1.2 (0.9–1.8)
Female	Reference	Reference	Reference
Birthplace			
Canada	19.7 (13.9–28.0)	19.8 (13.9–28.2)	16.3 (11.5–23.3)
Outside Canada	Reference	Reference	Reference
Disease Site			
Resp.	2.4 (1.5–4.0)	2.9 (1.7–4.9)	3.3 (1.9–5.7)
Non-Resp.	Reference	Reference	Reference
Resp. + Non-Resp.	1.8 (0.8–3.9)	2.4 (1.0–5.3)	2.3 (1.0–5.3)
Illicit drug use	3.8 (2.0–7.2)	4.4 (2.3–8.4)	3.3 (1.9–5.8)
Alcohol misuse	5.0 (2.8–8.9)	4.3 (2.5–7.7)	4.0 (2.4–6.8)

Abbreviations: SNVs, single nucleotide variants; OR, odds ratio; CI, confidence interval; aOR, adjusted odds ratio; Resp., respiratory; Non-Resp., non-respiratory.

^aAdjusted for age, gender, birthplace, disease site, illicit drug use and alcohol misuse.

8.3.4 Transmission occurs in both directions between Canadian-born and non-Canadian-born persons

Of the 93 genomic clusters in this study, 24 (25.8%) consisted entirely of CB persons, 26 (28.0%) entirely of nCB persons, 37 (39.8%) were heterogenous clusters with both CB and nCB persons, and six (6.5%) were unclassified, as they had a single individual of unknown birthplace amongst all CB or all nCB persons (**Figure 8-5**). Of these heterogenous clusters, 12 were predominantly CB with a median cluster size of 17 persons (interquartile range [IQR]: 10–51), eight were predominantly (>50%) nCB with a median of three persons/cluster (IQR: 3–4), and 17 were evenly split with 50% CB and 50% nCB persons. All but one of these evenly split clusters consisted of just two individuals—the remaining cluster was comprised of six individuals.

Transmission reconstruction within the heterogenous clusters, using a combination of WGS, epidemiological data, and review of case notes, suggested that transmission occurred in both directions. In 30 instances, a CB person was the most likely source of a nCB person's infection. Transmissions originating from nCB persons led to 31 TB diagnoses in CB individuals, although it should be noted that in seven of these transmission events, the nCB individuals transmitted a locally circulating strain they likely acquired from a CB person (**Figure 8-6**). The inferred setting of TB acquisition differed depending on the direction of transmission, with all but one of the 30 CB to nCB transmissions occurring within the community and 26 of these involving strains linked to large (≥ 10 cases) clusters. nCB to CB transmissions occurred from both household sources (29.0%) and the community (71.0%), and 77.4% involved strains belonging to small clusters (<10 cases/cluster), with 15 of these nCB to CB transmission events being two-person clusters.

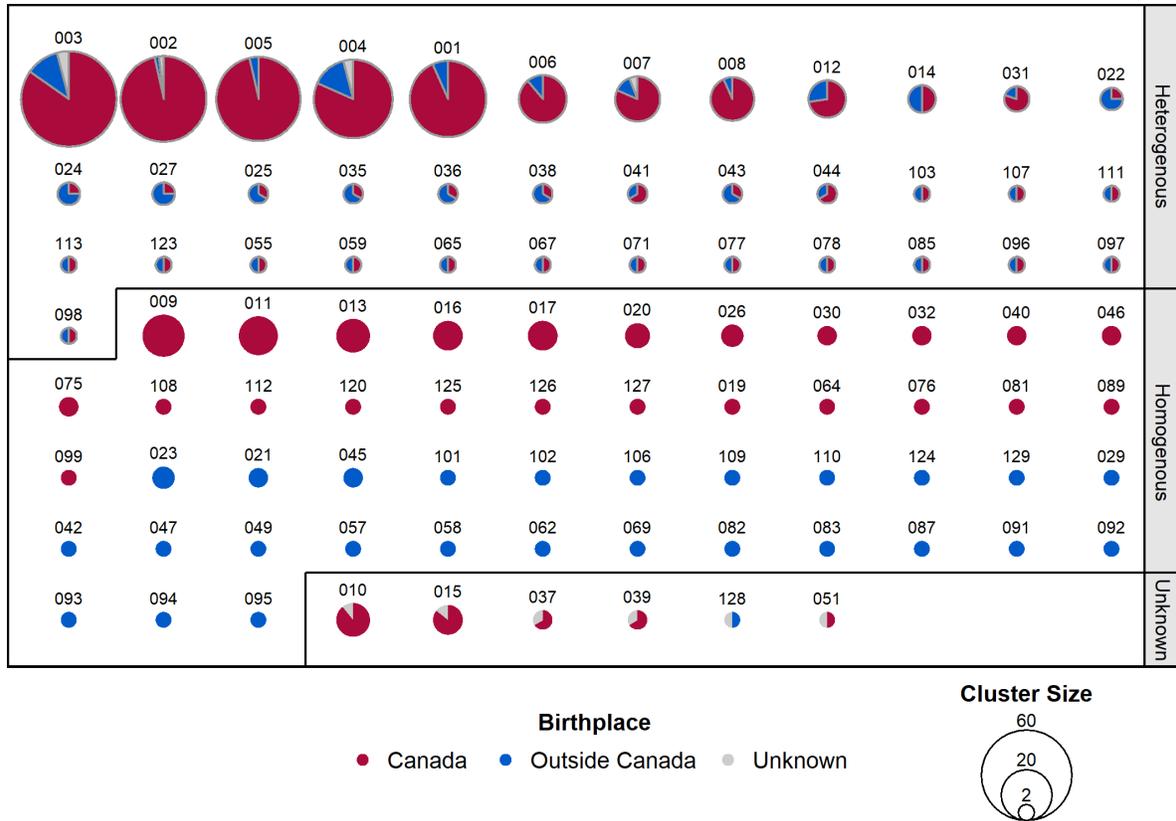


Figure 8-5. Genomic cluster sizes and birthplace composition. Pie charts representing all *Mycobacterium tuberculosis* genomic clusters ($n = 93$) based on a 20-single nucleotide variant threshold, British Columbia, 2005–2014. Categorized by clusters with a heterogenous birthplace, homogenous birthplace or clusters comprised solely of those born inside/outside Canada in addition to ≥ 1 case of unknown birthplace. Pie chart areas are scaled relative to the number of cases and the unique WClustID for genomic clusters are indicated above each pie.

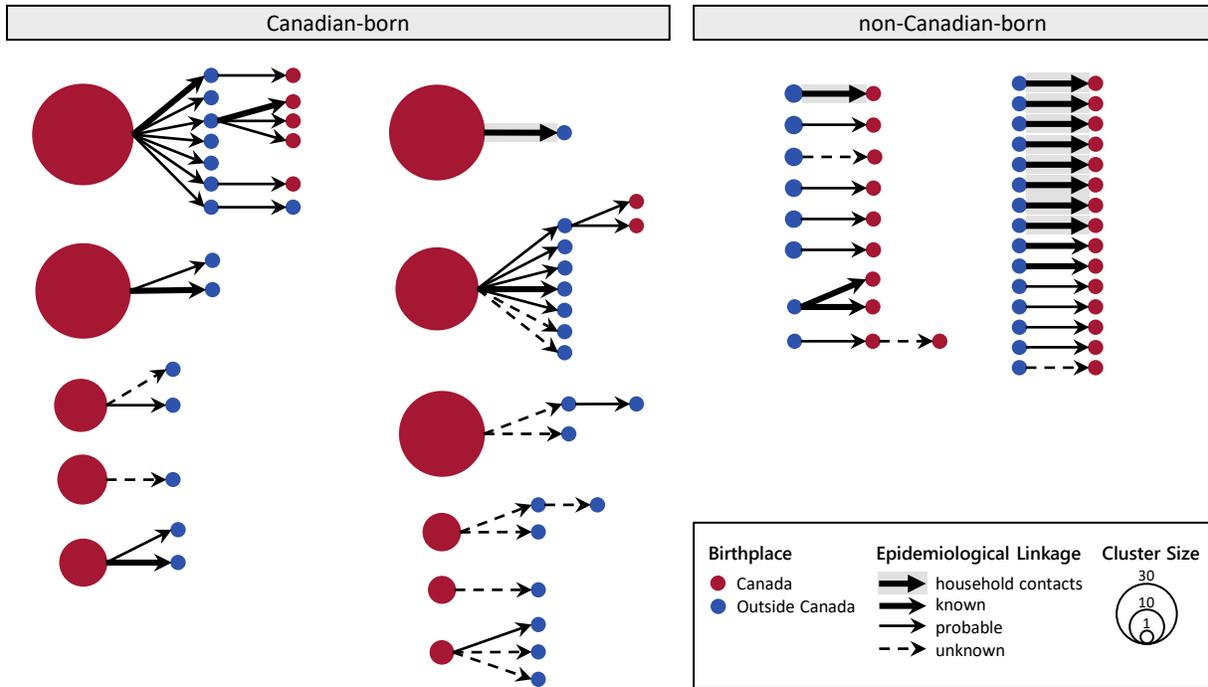


Figure 8-6. Mixed cluster transmission. Diagram depicting the number and direction of tuberculosis transmissions between Canadian-born and non-Canadian-born persons, according to genomic epidemiological investigations, British Columbia, 2005–2014. Transmissions are grouped by birthplace of initial source. Seven persons of unknown birthplace found within large clusters (≥ 10 persons) were not coloured separately. Circle size is proportional to the number of cases, and strength of the epidemiological evidence supporting linkages is indicated by different arrow type. Three small (2–3 persons) clusters were deemed not to represent local transmission and were excluded.

8.3.5 TB relapse vs reinfection

Accurate quantification of transmission also requires differentiating between TB relapse and exogenous reinfection; additionally, the true rate of relapse within a population is of interest to TB prevention and care programs, but is challenging to accurately measure with any technique besides WGS.⁴⁰⁰ Thirty-nine individuals diagnosed with TB during the study period had either a previous culture-positive TB episode in 2000–2004 or a second episode—and in one instance a third episode—of culture-confirmed TB within the study period, giving a total of 40 recurrence events. The genomic data for each episode’s isolate were examined relative to each other and within the context of the complete study cohort, and case notes for each individual were reviewed. It was determined that 32 (80%) recurrent episodes likely represented a relapse, while eight (20%) were new infections. Individuals deemed to have relapsed had *Mtb* isolates with a median of 1 SNV (IQR: 0–3) between episodes, whereas those with a new infection had a median of 205 SNVs (IQR: 36–503) (Figure 8-7). In the absence of WGS results for three reoccurrences, MIRU-VNTR genotyping was used instead to classify these episodes as a likely relapse ($n = 2$) or reinfection ($n = 1$).

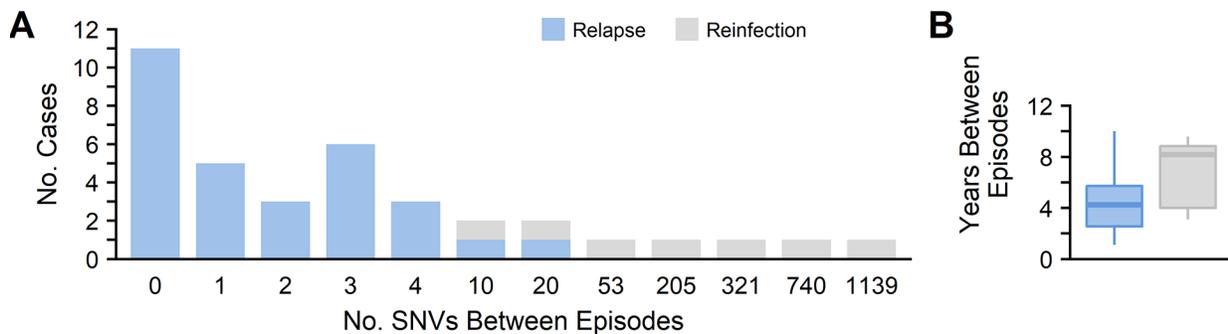


Figure 8-7. Recurrent tuberculosis characteristics. Differences in single nucleotide variants (SNVs) and time between episodes according to reoccurrence type, where the initial culture positive occurred Jan 1, 2000 or later and subsequent episode(s) during the study period (2005–2014), British Columbia, Canada. (A) Bar plot representing the number of single nucleotide variants (SNVs) observed between *Mycobacterium tuberculosis* isolates. (B) Boxplot representing the median and interquartile range for the number of years between episodes.

When the genomic profile of each recurrence isolate was examined against the background of the complete study cohort and episode timelines, it was not possible to establish a SNV cut-point to definitively distinguish between relapse and reinfection. Two cases had 10 SNVs between episodes, and, when viewed in the context of clinical, epidemiological, and physician narrative data, one was deemed to have relapsed while the other was likely infected with a new strain. Two cases had 20 SNVs between episodes, and similar to the individuals with 10 SNVs the data supported one as a relapse and one as a reinfection.

Of those individuals with a TB relapse, resistance to one or more first-line antibiotics was observed in the original isolate from seven individuals, one of which acquired resistance to an additional antibiotic by the time of the second episode. Six other isolates originally pan-sensitive, developed resistance to a first-line antibiotic by the time of relapse. It should be noted that nine of the relapse cases harboured strains common to their area of residence, thus it is possible that these individuals had been re-infected with a nearly identical strain to their original isolate. However, comorbidities and complications around treatment adherence for these individuals during their first episode suggested that relapse rather than re-infection was most likely.

Among the eight individuals with a reinfection, three were CB, with both episodes involving strains known to circulate in their communities, and all three had substance use as a risk factor for TB infection. The remaining five were nCB persons and their isolates were unique within the BC collection, indicating that these individuals likely acquired their infections outside BC. Three of these individuals were known to have travelled to their high-incidence TB birth country between episodes.

8.3.6 Characterization of large genomic clusters

To understand the patterns driving local transmission beyond the household contact level, the characteristics of large (≥ 10 persons) genomic clusters were examined and found that all were comprised of predominantly CB persons, with 8 of 11 large clusters concentrated in metro regions (**Table 8-4**). The median age of individuals varied from 40 to 58 years across the clusters ($p = 0.001$), with the youngest cluster (WClust-5) also representing the single cluster comprising more females than males (57.1% versus 42.9%). While the proportion of individuals with one or more risk factor for TB (HIV, illicit drug use or alcohol misuse) varied between clusters, 76.8% (255 of 332 with known risk factor data) of persons in a large genomic cluster had at least one risk factor. With the exception of the two smallest clusters, under-housed persons represented $>25\%$ of individuals in any large cluster. Clusters with individuals residing in predominantly rural areas had the lowest proportion of under-housed persons.

With respect to epidemiological linkages between individuals within the same genomic cluster, the number of persons with known connections (i.e. named contacts) varied substantially (0.0–83.9%), with the three largest clusters (WClust-3, WClust-2, WClust-5) having the largest proportion of cases with known epidemiological linkages. WClust-5 was particularly well-characterized, with 47 of 56 persons (83.9%) connected to another case within the cluster. Among the nine persons not linked to another case, all had TB risk factors, two were under-housed and five were noted as contacts of active TB cases in the case narratives; however, the names of these individuals were not specified. Eight of the nine were diagnosed in the geographic region in which WClust-5 is concentrated and, while the remaining individual was diagnosed outside this region, they previously resided there.

Table 8-4. Large genomic clusters. Characteristics of genomic clusters (20 single nucleotide variant threshold) comprised of ≥ 10 individuals, British Columbia, 2005–2014.

Cluster ID	Cluster Size	Canadian-born ^a <i>n</i> (%)	Median Age (IQR) years	Gender M:F	Risk Factors ^b <i>n</i> (%)	Under-housed ^c <i>n</i> (%)	Predominant Community Type (%)	Known epi-links ^d <i>n</i> (%)
WClust-3	72	61 (88.4)	50 (43–57)	11.0	49 (74.2)	47 (65.3)	Metro (76.4)	29 (40.3)
WClust-2	56	54 (96.4)	46 (36–54)	1.4	36 (78.3)	15 (26.8)	Rural (51.8)	37 (66.1)
WClust-5	56	54 (96.4)	40 (29–48)	0.8	42 (80.8)	22 (33.9)	Rural (75.0)	47 (83.9)
WClust-4	49	40 (85.1)	49 (39–54)	3.9	33 (71.7)	23 (46.9)	Metro (79.6)	9 (18.4)
WClust-1	46	43 (93.5)	44 (35–54)	1.6	37 (84.1)	31 (67.4)	Metro (87.0)	16 (34.8)
WClust-6	18	16 (88.9)	42 (33–60)	1.8 ^e	11 (64.7)	8 (44.4)	Metro (83.3)	2 (11.1)
WClust-7	16	13 (86.7)	50 (48–56)	3.0	13 (86.7)	11 (68.8)	Metro (93.8)	6 (37.5)
WClust-8	15	14 (93.3)	44 (40–54)	2.0	12 (85.7)	7 (46.7)	Metro (80.0)	2 (13.3)
WClust-9	14	14 (100.0)	48 (44–58)	1.0	10 (83.3)	10 (71.4)	Metro (57.1)	0 (0.0)
WClust-11	12	12 (100.0)	58 (44–70)	3.0	6 (66.7)	1 (8.3)	Rural (83.3)	3 (25.0)
WClust-12	11	8 (72.7)	51 (38–58)	2.7	6 (54.5)	1 (9.1)	Metro (72.7)	4 (36.4)

^aData not available for 7 individuals, percentage represents those with complete data.

^bOne or more risk factors (HIV, illicit drug use, or alcohol misuse); data unavailable ($n = 33$), percentage represents those with complete data.

^cPersons with no fixed address or living in a shelter, group home, or single-room occupancy housing.

^dKnown epi-links (epidemiological linkages) represent individuals that were named contacts of another within the cluster.

^eOne transgender/gender-unknown individual excluded from analysis.

Previously, genomic reconstructions of two of the largest clusters were described in the literature—these were pilot studies that enabled BCCDC TB Services to obtain funding for the present dissertation work. Here, the timeline (**Figure 8-8**) and likely transmission pathways for another of the largest genomic clusters (WClust-2) was reconstructed (**Figure 8-9**). This cluster comprises individuals residing in eight BC communities spanning a maximum distance of 1000km, with cases largely concentrated in the northern region of the province. Many cases resided in rural (51.8%) or remote (5.4%) areas, with 35.7% of cases residing in urban/rural areas, and 7.1% residing in a metro area. It was found that 53.6% (30/56) of isolates had zero SNVs when compared to at least one other isolate and 89.3% (50/56) were within five SNVs of another isolate. However; when examining the individual SNV profiles, it appears that TB in at least 18 individuals was likely the result of initial transmission prior to the study period. In one instance, both the epidemiological and genomic evidence support this—a case who relapsed within the study period and who had their first episode several years prior was identified as the source of at least one known transmission and one possible transmission event during this first episode.

Among the 56 cases, 37 individuals identified one another within WClust-2 as a contact—this is not surprising, given the rural and remote nature of the population; however, only 12 of these named individuals were plausible source cases, as revealed by the SNV profiles. The remaining six individuals either acquired their infection from a different, unnamed individual in the cluster or they represent reactivation of LTBI acquired before the study period. In five cases, individuals within WClust-2 named contacts not belonging to WClust-2 who instead harboured genomically distant isolates from strains circulating in the same region.

One individual thought to be part of WClust-2 was diagnosed on autopsy (indicated as case “x” in **Figure 8-9**) and an isolate was not available for genotyping or WGS. This individual had respiratory TB and was believed to be highly infectious—epidemiological evidence supports this individual as the likely source for three subsequent culture-confirmed cases. Also of interest in this cluster is the high number of individuals that represent super-spreaders—transmitting to multiple persons that went on to develop culture-confirmed TB. Of note, isolate #21 represents

an individual that presented to healthcare on multiple occasions with signs and symptoms of tuberculosis, and was hospitalized for several days without consideration of TB as a cause of their respiratory infection. This individual was not diagnosed until five months later and was potentially the source for at least four secondary cases, and ultimately a transmission chain involving seven culture-confirmed cases.

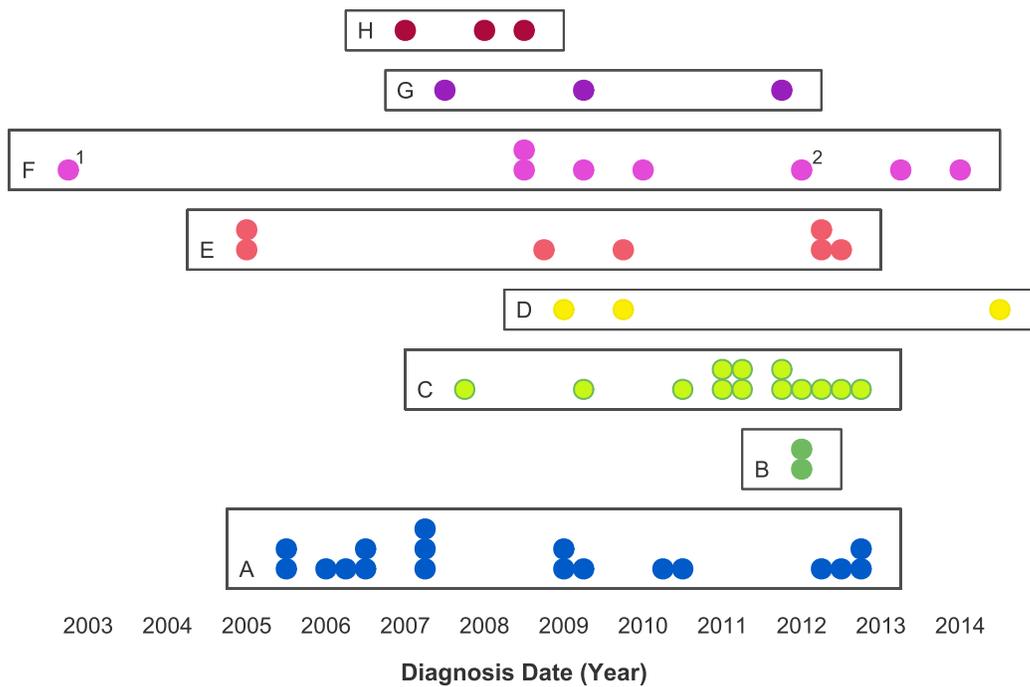


Figure 8-8. Timeline of case diagnosis. Number of tuberculosis cases organized by year and quarter of diagnosis over a 10-year period in British Columbia, Canada. Each circle represents a single case, and are coloured by community (A – H) in which the individuals reside. The subscript numbers “1” and “2” indicate the first and second episode for the relapse case.

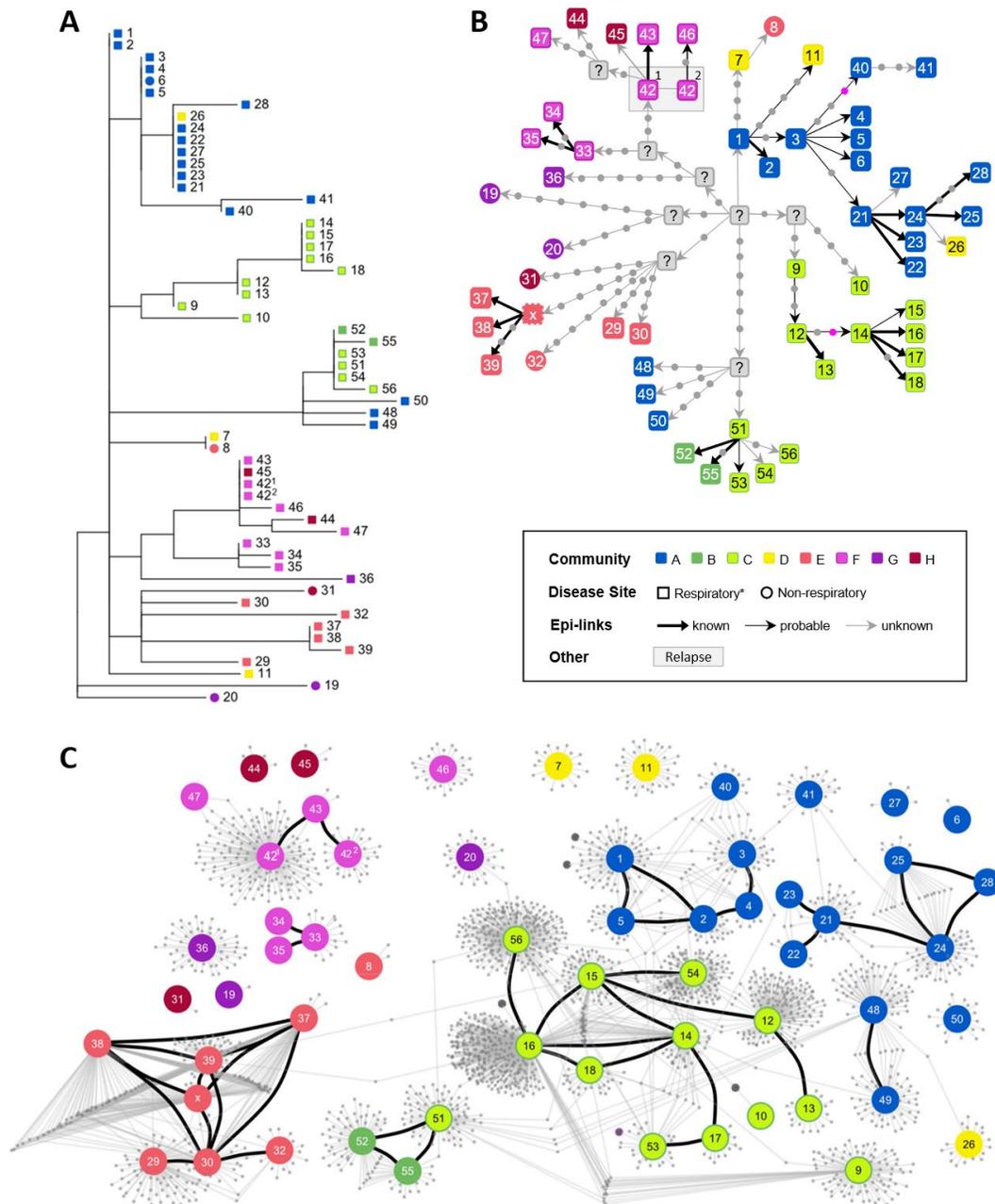


Figure 8-9. Characterization of WClust-2. (A) Phylogenetic tree. (B) Putative transmission pathways. Each node represents a single culture-positive isolate (with the exception of case “x”, diagnosed on autopsy and for which no *Mtb* isolate or genetic material was available). Nodes labelled with “?” represent putative source cases outside the study period. Arrows indicate a transmission event, with the arrow type indicating the strength of the epidemiological connection between individuals. The number of single nucleotide variants (SNVs) between isolates are indicated along the arrows as small circles—grey for fixed SNVs and pink indicating the presence of a minority variant. One case, enclosed in a grey box, represents a relapse where the second episode fell within the study period, thus both isolates were included in the reconstruction. *Excludes “other respiratory” sites. (C) Network of named contacts. Medium dark grey circles represent persons with culture-positive isolates genomically distant from WClust-2. Small grey circles represent persons named contacts of individuals in WClust-2 in which culture-positive TB was not diagnosed. The subscript numbers “1” and “2” indicate the first and second episode for the relapse case.

8.4 Discussion

In this ten-year retrospective study, 2,303 *Mtb* isolates were successfully genotyped and the genomes of 1,221 were sequenced—including all 974 genotypically clustered and 247 special interest isolates. WGS indicated that the proportion of culture-confirmed TB cases arising from local transmission in BC is 21.8% using the “ $n - 1$ ” method⁸⁸ and a 20-SNV threshold. This is a much more accurate quantification versus previous estimates made using MIRU-VNTR data alone.

The most important risk factor for belonging to a genomic cluster was Canadian birth. In contrast, the majority (75.8%) of genotypically clustered nCB persons were no longer clustered after genome sequencing. This is an important observation, revealing the shortcomings of MIRU-VNTR as a tool for understanding transmission and identifying outbreaks.^{121,401} Many of these predominantly nCB genotypic clusters instead likely represent an *Mtb* genotype common to individuals who have emigrated from the same region of the world—an observation that has been reported in other low-incidence TB settings.^{336,354} The ability to distinguish between LTBI reactivation and local transmission is key to developing evidence-based TB prevention strategies, and the quantification of cases attributable to local versus overseas acquisition allows TB prevention programs to appropriately direct their interventions, whether they be enhanced contact investigations to prevent local transmission or better screening and LTBI treatment programs for migrants. Beyond Canadian birth, illicit drug use and alcohol misuse were identified as significant risk factors for clustering. Indeed, most individuals in large genomic clusters had one or more of these risk factors, and often reported a lack of adequate housing, highlighting the need for TB prevention programs to focus on vulnerable populations to curb ongoing local spread of TB.

While transmission does occur between CB and nCB populations, it is not very common. Not surprisingly, transmission from nCB persons most often represented single transmission events within households—a number of which were nCB adult family members transmitting to CB children, as described in **Chapter 5**. Although stories in the media often state that immigrants pose a TB risk to the population of the country to which they are immigrating,^{402,403} the results

reported here support the findings of other studies which have repeatedly shown that nCB persons do not pose a substantial TB risk to CB persons, as transmission from these individuals is rare.^{404,405}

Clustering rates varied significantly between lineages, with the highest level of clustering amongst Euro-American isolates. This is in due in part to the strong phylogeography of TB and Canada's history—early European settlers introduced this lineage to the population.⁹² Additionally, studies have indicated that outbreaks occur more often with Euro-American and East-Asian lineage strains as a result of phenotypic differences in the virulence of these lineages compared to the others.^{44,45} Indeed, the only other lineage representing a large genomic cluster within BC was the East-Asian lineage (WClust-6). This cluster was comprised of predominantly CB persons, including the earliest cases, suggesting that this lineage was likely introduced to BC and was circulating locally prior to the study period.

With respect to TB reoccurrences, WGS can help to distinguish between a relapse or reinfection, more so than MIRU-VNTR; however, where individuals are exposed to a common strain circulating in their community after treatment of their initial episode, it may not always be possible to differentiate between a relapse or new infection. In this study, there were several cases for which the presumed relapse isolate was 0–2 SNVs from another individual's isolate and could have represented reinfection, although in these instances, it was determined that these were most likely relapses due to comorbidities and treatment compliance issues increasing the likelihood of treatment failure.

Reconstructing the putative transmission pathways in a large genomic cluster revealed that contrary to the assumption of a single outbreak, WClust-2 represents multiple discrete chains of transmission with strains diverged from a shared common ancestor circulating in the northern region of BC. Reactivation of a previously acquired infection in a single individual is frequently the seed for a new localized outbreak, with super-spreaders—often individuals that for various reasons⁴⁰⁶ experience a diagnostic delay and thus increase the risk of TB spread^{6,406}—driving much of the transmission in these large clusters. While many features, such as super-spreading

and transmission amongst marginalized populations, are shared across the large clusters, each cluster differed slightly based on the community, sub-population, and geography of the region. In contrast to WClust-2, the large clusters WClust-3 and WClust-5, which have been detailed previously,^{74,173} were highly concentrated in particular populations and regions, with one linked to specific homeless shelters and the other illicit drug use.

The findings presented here highlight the importance of conducting population-based retrospective studies to build a knowledge base of genomic clusters and provide context for isolates that are prospectively sequenced. While this work was retrospective, access to genomic data at the time of investigation is key to quick and accurate identification of significant clusters as well as those individuals within a cluster contributing the most to onward transmission. Epidemiological information alone is insufficient—within several clusters, individuals named contacts who, after sequencing, were determined to belong to different genomic clusters. Likewise, genome sequencing alone is often insufficient to fully derive transmission networks.

While this large, retrospective study over a ten-year period represents a comprehensive investigation of the genomic epidemiology of TB in BC, there are several limitations to note. First, not all 2,303 genotyped isolates were sequenced, only those which that were assumed to have been locally transmitted based on MIRU-VNTR data or other epidemiological information. Thus, it is possible that this sequencing strategy may have missed isolates that were indeed part of a local transmission cluster, just as it missed culture-negative cases or those for which no genetic material was available.

The study was also limited by the retrospective nature of the epidemiological data collected, with no ability to ask additional questions of individuals that were prompted by the genomic findings, such as connections to a particular location/region. Genomics-guided investigations, with questions tailored to a specific cluster, may have uncovered more epidemiological linkages between individuals than classic contact investigation techniques, particularly when the transmission event leading to disease occurred several years prior—contact investigations are generally focused on more recent exposures, and failure to ask about earlier potential sources of

TB may lead to missed connections within the transmission network.^{6,68} Accurate source case identification is also hampered by *Mtb*'s long latency period, and even with a ten-year study window, there were still challenges in identifying source cases.

In conclusion, nearly 20 years after the World Health Organization declared TB a global health emergency, the disease still affects 10 million people per year and was responsible for 1.7 million deaths in 2016.⁵ Innovative new tools are necessary to both interrupt current transmission and reduce the pool of infected individuals who might reactivate in the future, seeding an outbreak. This study demonstrated the value of WGS for large population-based studies, providing a benchmark transmission estimate against which future progress towards elimination within BC's TB program can be compared. It is clear that WGS offers significant value over traditional genotyping, and is able to address both population surveillance and case investigation needs. Routine use of WGS will significantly improve our understanding of TB transmission, and provide the evidence necessary to develop more effective care and treatment strategies as we move towards TB elimination goals.

Chapter 9: Conclusion

9.1 Summary of findings

The overarching aim of this dissertation was to describe the molecular and genomic epidemiology of TB in British Columbia with a view to increase our overall understanding of local TB transmission and identify risk factors related to person-to-person spread with a particular focus on large clusters. In doing so, the findings presented here have added to the evidence base for public health follow-up of TB cases. A summary of the major findings of the dissertation are listed below.

In **Chapter 2**, a comparison of the previously established on-request and new universal genotyping program revealed that where requestors specified an individual or outbreak comparator, they had only slightly better than a 50% chance of correctly matching the individual or strain. Furthermore, few genotypic clusters had all isolates requested—including several large clusters of public health concern. The findings underscore that the social networks of persons with TB are complex and incompletely understood, with unknown epidemiological linkages between individuals. They also demonstrate that a universal approach is superior to a limited, on-request genotyping program, and suggest that equipping TB personnel with molecular epidemiology clustering from a universal genotyping program will allow TB programs to better focus and prioritize investigations.

Extending universal genotyping to a ten-year period, **Chapter 3** details a comprehensive analysis of the molecular epidemiology of TB in BC. More than 1,500 distinct *Mtb* genotypes were detected across four major *Mtb* lineages, reflecting BC's diverse population. Based on genotype clustering, it was estimated that over one-third of BC's TB cases were potentially the result of local transmission. Large clusters of predominantly Canadian-born persons were identified and analyses of clinical, epidemiological, and demographic characteristics revealed these to be discrete populations with distinct disease and risk factor profiles, likely representing true local transmission, and ideal groups to which to target specific interventions.

To place the findings of the previous chapter in the larger context of TB molecular epidemiology elsewhere in Canada, an analysis was undertaken in which the genotyping data and selected clinical and demographic data were compared across BC and Ontario (**Chapter 4**). Among 4,916 isolates, there were 3,461 unique genotypes. Of these 175, were common to both BC and Ontario, which represented just 18.0% of Ontario's and 31.6% of BC's isolates. Interestingly, several large genotype clusters known to represent transmission among predominantly Canadian-born persons within each respective province were not observed in the other province. Genotypes common to both provinces were largely associated with persons born outside Canada and representing infections acquired elsewhere. These findings underscore the importance of understanding regional epidemiology in order to interpret interprovincial genotypic clustering and indicate that interprovincial transmission events, at least in geographically distant, large, immigrant-receiving provinces, are uncommon.

Through a combination of genotype and genomic analyses, in **Chapter 5** it was determined that nearly a third of pediatric TB infections in BC were the result of local transmission from an adult—most often a non-Canadian-born family member for children of immigrant parents or an adult contact from the community for Canadian-born children of Canadian-born parents. Importantly, travel to the high-burden TB countries of their parent's birth was the source of infection for at least 50% of children born in Canada to immigrant parents. These findings reveal the role that age, birthplace, and parents' birthplace play in determining a child's likely route of TB exposure, and stand in contrast to the assumption that most pediatric TB cases arise from household transmission.

Building on the knowledge gained through the interprovincial comparison of *Mtb* genotypes, the next analysis used genomics to explore TB transmission within Yukon Territory (YT) and between YT and BC. **Chapter 6** details the TB transmission dynamics within YT and characterizes three large clusters, each with a super-spreader—an individual who was likely the source for multiple secondary cases. The finding that transmission also occurs between YT and BC residents highlights the need for routine molecular-based surveillance and communication across jurisdictions with known frequent cross-border movement.

Taking advantage of the long-standing relationship with the community and extensive knowledge YT nurses have of the small, well-defined population of persons with TB, a mixed methods study was conducted to both compare traditional field epidemiology to molecular-based approaches, and understand the value added by using molecular and genomic methods in a remote setting with well-described epidemiology (**Chapter 7**). Both field epidemiology and molecular methods correctly assigned individuals to specific clusters; however, neither method on its own proved capable of regularly identifying source cases and specific transmission events. Together, genomics and detailed field epidemiology yielded the highest-resolution insights into transmission and could be used to guide future contact investigations. Participant feedback indicated that receiving WGS data was preferable to genotyping results, and that in-person training in how to interpret WGS results was desired by the YT TB prevention team.

Chapter 8 describes the genomic epidemiology of tuberculosis in BC using WGS data of *Mtb* isolates, linked to key clinical, epidemiological, and demographic information. Genomics refined the previous estimate of recent transmission rates in BC, revealing that approximately 21.8% of TB infections were locally acquired. Genomics-based investigation of the largest clusters—all of which predominantly comprised Canadian-born persons—identified distinct sub-populations and risk factors for transmission. Although each of these clusters had some unique characteristics, common factors associated with belonging to a large genomically-defined cluster were substance use and Canadian birth. Notably, Canadian-born persons had considerably higher odds of belonging to a cluster (aOR 19.7, 95%CI: 13.9–28.0). In contrast, among the small proportion (7.7%) of persons born outside Canada that acquired TB locally, no large genomic clusters were seen. Reconstruction of putative transmission pathways for a large cluster revealed several smaller sub-clusters associated with specific communities and genomic variation—indicating this strain has likely been circulating in BC’s northern region prior to the study period. The transmission patterns observed in this cluster stand in contrast to the two large BC clusters described previously in the literature^{74,173}—outbreaks concentrated in a specific location/region with key super-spreading individuals—and provide additional insight into the drivers of transmission within BC.

To summarize, the use of molecular-based methods, particularly whole genome sequencing, in combination with detailed epidemiological data has revealed deeper insights into BC's *Mtb* population structure and transmission dynamics than has ever before been possible. The quantification of the proportion of TB attributable to local transmission, and the identification of specific risk factors that play an important role in the ongoing person-to-person spread of TB within the province, will ultimately inform improved monitoring, surveillance, and resource allocation and inform the development of new prevention strategies.

9.2 Unique contributions, implications and impact

The series of studies comprising this dissertation have contributed greatly to our understanding of the genomic epidemiology of tuberculosis in BC and together they represent the largest genotyping and genomic investigation of TB ever undertaken in North America. Nearly all culture-positive TB cases diagnosed in BC from 2005 through 2014 ($n = 2,290$) were genotyped using 24-locus MIRU-VNTR, and 1,221 of these isolates were submitted for WGS—including all genotypically clustered isolates ($n = 974$). Through this analysis, it was possible to more reliably distinguish between cases arising from LTBI reactivation versus local transmission, and has provided a better estimate of transmission than has ever been possible.

Analysis of the ten-year BC cohort revealed that the likelihood of clustering, particularly in large clusters (≥ 10 persons), was higher in rural settings compared to metropolitan areas. Previous studies of TB clustering elsewhere have largely focused on urban areas, and thus this finding's implications are critical to appropriately directing TB prevention and care resources to these often underserved communities and to educating clinicians and public health personnel in these areas to better recognize TB cases in order to rapidly diagnose and treat. An equally important finding was the determination that at least half of pediatric TB cases born in Canada to immigrant parents most likely acquired their infection through travel to the high-burden TB countries of their parent's birth. This has implications for pediatric TB investigations, but more importantly extends to policy around individuals born outside Canada, who are subject to TB screening during immigration but not upon returning to Canada after travelling to their country

of origin. The data suggest that new recommendations around screening individuals with recent community-based travel to high-incidence settings may be warranted.

While several genomics studies of outbreaks in the remote regions of Canada's North have been conducted,^{166,167} the Yukon studies presented in this dissertation are a first. The genomic epidemiology of YT revealed considerably different TB transmission patterns relative to other northern regions, likely due to the different history, community structure, and demographics of YT versus other northern territories. Also, in part due to frequent travel between BC and YT, which, unlike other northern settings, has year-round road access. WGS provided definitive proof that transmission does indeed occur between YT and BC residents and highlights the need for routine molecular-based surveillance and communication across these two jurisdictions.

The genomic epidemiology approach used in these studies allowed for the identification of linkages that traditional contact investigations failed to detect. This has provided important data for understanding the sub-populations and risk factors associated with TB transmission in BC and YT. Particularly informative was the large cluster reconstruction using WGS, which has the ability to determine the direction of transmission—not possible with classic genotyping techniques. Genomics identified clusters and revealed many individual transmission events within each cluster, giving an indication of the role that super-spreaders play in transmission and a deeper understanding of the characteristics common to persons who spread TB to multiple other individuals. These insights will have a direct impact on TB prevention policies and guidelines, and lead to the development of improved contact investigation methods supported by genomics-based approaches.

Findings from this dissertation add to the growing body of literature investigating the use of WGS for diagnostics and clinical practice. At this time much of the focus remains on the use of WGS for research purposes or in individual outbreak investigations, and few mycobacteriology laboratories in the world have rolled out WGS for all culture-positive TB isolates. The large, comprehensive retrospective population-based study described herein serves as a demonstration of the powerful insights attainable through universal WGS and will provide critical guide as BC

considers implementing routine WGS. If the province implements WGS for routine use, these data suggest that there will be subsets of culture-positive TB cases in which WGS is likely to provide actionable insights for contact investigation and outbreak management. The data also provide a retrospective data bank against which isolates sequenced in the future can be mapped—given TB’s potentially long latent period, having this historical database to back-compare against is important, and that is something that to the best of my knowledge no other public health agency currently has.

9.3 Strengths and limitations

This dissertation’s scope—no other study has used genotyping and genomics to look at as large and complete a retrospective TB dataset as this, its uniqueness—few jurisdictions have access to linked clinical, epidemiological, and demographic data as well as the capacity for WGS, and the extensibility of the findings to other TB programs in low-incidence settings all represent major strengths of this work. Sampling 99.3% of culture-positive *Mtb* isolates over a ten-year period reduced the potential for bias and provided validity to the conclusions. The comparison of *Mtb* genotypes with Ontario data placed the BC results in the larger Canadian context and confirmed the molecular epidemiology study conclusions in **Chapter 3**. The ability to work with smaller datasets with detailed epidemiological case-level information, such as the pediatric and Yukon Territory cohorts, permitted the testing of the approaches to transmission reconstruction, facilitating the later work on the large WGS study (**Chapter 8**), while also providing interesting and relevant results. Additionally, all eight research studies included in this dissertation received input from clinical, field-based TB program officers—these co-authors critically revised each manuscript and provided important insight into how the findings could impact TB policy and practice.

Nevertheless, there are several limitations that should be noted, including the important point that much of the case-level data used throughout this dissertation were collected for purposes other than to test specific research hypotheses; for example, many data fields were collected for broad, population-level surveillance activities, rather than individual transmission inference and contact investigation. Furthermore, all studies were retrospective in nature—meaning the

genotyping and WGS results were not available at the time of investigation and there was no opportunity to confirm molecular-based linkages that had not already been captured through the original case interviews.

An inherent limitation of all TB genotyping and genomics studies is the current requirement for large volumes of input DNA. This means that genotyping and sequencing must begin with an *Mtb* culture, which restricts the study population to only culture-confirmed TB cases. This may result in missed connections in a transmission chain. However, individuals with culture-negative TB (~20% in BC) are generally considered non-infectious and are unlikely to contribute to the spread of the disease.⁴⁰⁷ Other possible reasons for missing links within chains of transmission are instances in which an infectious individual left the province or died prior to diagnosis. It is difficult to estimate how often this may have occurred, but the contribution of these missed cases should be considered in future studies, particularly when investigating TB transmission amongst Canadian-born, street-involved individuals. Given the current opioid overdose crisis in BC and the association of TB and substance use, it is reasonable to expect that amongst the province's annual opioid deaths—over 1,400 in 2017⁴⁰⁸—there may be some active TB cases. Under the BC Coroners Act, coroners have the authority to authorize a post-mortem examination if the coroner deems it necessary for the purposes of the investigation. From 2015–2017, a post-mortem examination was conducted in under a quarter of suspected illicit drug overdose deaths (Andrew Tu, BC Office of the Chief Coroner, personal communication).

While WGS has provided significant insights into TB transmission, there remain several challenges in using WGS data as part of routine TB prevention and care, even in a well-resourced setting such as BC. First, the costs of large-scale, routine WGS remain considerable, despite the significant reductions in sequencing prices seen over the past decade. Culture and sample preparation still incur substantial reagent and technical personnel costs, and sequencing economy is best achieved when batching hundreds of isolates per run—sequencing BC's infrequent TB cases as they arise is lower-throughput and higher cost. To mitigate these costs, WGS for this dissertation focused only on the subset of genotypically identical isolates, along with isolates closely related to known transmission clusters by MIRU-VNTR, e.g. one- or two-

locus mismatches. Although the sequenced isolates likely captured nearly all the local transmission events, it is possible that a small number of transmitted cases were not sequenced due to larger-than-expected differences in genotype results. Second, 10% of isolates were genomically indistinguishable from one or more isolates in the study population, and in some cases, even the addition of contact tracing data did not suggest specific individual transmission events within these clusters. Lastly, interpretation of WGS results is complicated by the currently available laboratory methods. Sequence heterogeneity may be introduced by laboratory culture of isolates—an issue that will be resolved once we can reliably sequence directly from specimens³⁷³—or may occur naturally as a result of within-host microevolution.^{162,409} Both issues could lead to erroneously excluding individuals from a transmission chain. While it is quite possible these issues may have occurred in this study, the large number of individuals with no genomic variation over a number of years and multiple hosts, some with long infectious periods, indicates that this may not be a widespread phenomenon.

In the early chapters of this dissertation, a number of limitations were raised which were subsequently addressed in the ensuing studies, particularly with respect to the overestimates of clustering using MIRU-VNTR genotyping, which WGS largely resolved. However, the concerns raised around classifying all persons born outside Canada as a single group—described in **Chapter 3**—had been left unanswered. Therefore, a small analysis to investigate this potential limitation was undertaken, the findings of which are detailed in the next section.

9.3.1 Homogeneous Group or a Multicultural Mosaic? The Challenge with Reporting Birth Outside Canada as a Tuberculosis Risk

Background

Birth in a country with high tuberculosis (TB) incidence—greater than 30 TB cases per 100,000 population⁶—or very high incidence (≥ 200 cases/100,000) is a major risk factor for TB disease in many low- (<10 cases/100,000 population)⁴¹⁰ and medium-incidence (11–29 cases/100,000) settings, where TB rates are largely driven by reactivation of latent tuberculosis infection (LTBI) acquired in an individual’s country of birth.⁴¹⁰ Thus, population-level TB surveillance programs and research studies typically collect birth country for each individual. However, these data are often reported in aggregate—all persons born outside Canada are reported as a single group, without stratification based on birth country or age at immigration. In these studies, a person who immigrated from a low-incidence country as an infant and an adult recently arrived from a high-incidence country would be similarly classified, despite likely having different TB exposure histories. Like many low-incidence countries, Canada is multicultural, welcoming over 250,000 permanent residents from ~200 countries in 2015.⁴¹¹ Migrants arrive as both temporary and permanent residents and include refugees, students, and skilled workers.⁴¹² More than 60% of new permanent residents come from just 10 countries, where TB incidence ranges from low (USA, France, UK) to medium (Iran, Syria) to high (China) and very high (Philippines, India, China, Pakistan, Nigeria).⁴¹¹ Reporting these populations under a single label—typically “foreign-born”—limits our understanding of local TB epidemiology, and may stigmatize immigrants, particularly around public understanding of TB transmission within Canada and risk to the population. Here, it is examined how characterizing migrants based on TB incidence in their birthplace improves our understanding of regional epidemiology and transmission.

Methods

Two previously described cohorts were used—the **Chapter 3** study population and cohort detailed by Ronald et al. (2018), to examine how stratifying persons born outside Canada according to TB rates in their country of birth affects the understanding of TB trends. With the first cohort, individual-level clinical and demographic data was extracted from the British Columbia (BC) Provincial TB registry⁴¹³ and linked these to genotypic data representing 99.3% of all culture-positive TB cases in BC, 2005–2014, as previously described (**Chapter 3**). Separately, in a second cohort, records from the BC Provincial TB Registry⁴¹³ were linked to data from the Immigration, Refugees and Citizenship Canada Permanent Residents database and Population Data BC health administrative databases^{414,415} to calculate TB incidence rates and time from immigration to active TB as described previously.⁴¹⁶ Birthplace TB incidence rates were derived from yearly country-level incidence data (all forms of TB/100,000 population).⁷⁹ For the genotypic analysis, the cohort was divided into two groups—individuals from high-to-very-high-incidence settings and those from medium-to-low-incidence settings. For the population-based TB rate analyses, three groups were used: very high incidence, high incidence and medium-to-low incidence.

TB rates according to years following immigration and case counts by TB incidence in country of birth were represented graphically, with the time trends in case counts evaluated by linear regression. Demographics and genotypic clustering of individuals between groups was compared using summary statistics including median, interquartile range (IQR) and t-test, where appropriate. Analyses were conducted in R (v3.4.1).

Results

Amongst 2005–2014 culture-confirmed TB diagnoses ($n = 2,290$), the number of cases was lowest in persons born outside Canada in regions with medium-to-low TB incidence (**Figure 9-1**). While case counts were stable in this group ($p = 0.522$) and the Canadian-born ($p = 0.636$) after 2010, they increased in persons from high-incidence countries ($p = 0.005$). TB incidence rates among permanent residents arriving in BC between 2005–2012 reflect a similar pattern—the overall rate in persons from very high-incidence countries (53.6/100,000 person-years) was

>20-fold higher than in those born in a low-incidence country (2.6/100,000 person-years). Furthermore, there is a clear trend whereby TB incidence rates after arrival in BC are highest among residents from very high TB incidence countries (**Figure 9-2**). Age at time of diagnosis, and years between arrival and TB diagnosis also vary substantially amongst those born outside Canada, from a median age-at-diagnosis of 53 years (IQR: 35–73) and 12 years (IQR: 3–21) since arrival for persons born in high-incidence countries, to 66 years (IQR: 48–78) and 36 years (IQR: 14–50) for those born in medium-to-low-incidence countries ($p < 0.001$).

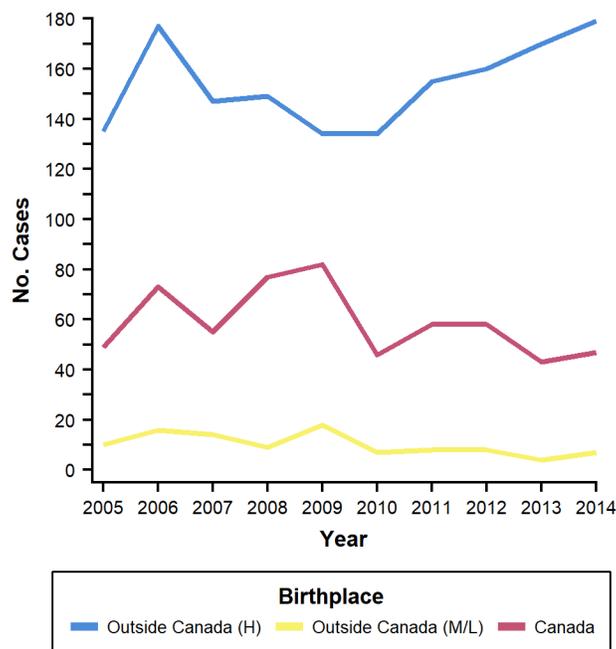


Figure 9-1. Trends in active tuberculosis diagnoses in British Columbia (BC), Canada. Number of culture-confirmed cases over a ten-year period categorized by birthplace: Outside Canada (H—high-incidence countries, M/L— medium to low-incidence countries), Canada.

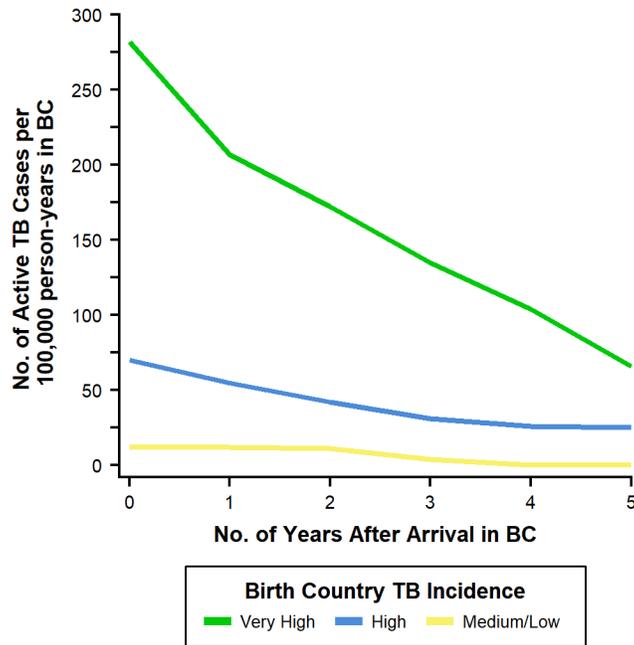


Figure 9-2. Tuberculosis incidence rates in British Columbia (BC) for persons born outside of Canada. Plot represents TB cases who immigrated to BC from 2005 through 2012, stratified by country of birth incidence (very high, high, and medium to low-incidence) and number of years in BC.

TB Transmission Within BC Varies According to Birthplace

Next, genotypic clustering was examined in the linked cohort. Of the 1,641 persons born outside Canada in the dataset of 2,290 genotyped isolates, 69.8% ($n = 1,146$) grew a TB isolate with a unique genotype, suggestive of LTBI reactivation. 1,071 (93.5%) of these 1,146 cases emigrated from high-incidence settings. Examining persons born outside Canada in the study, two distinct groups were observed. First, in large clusters (≥ 10 cases) consisting of predominantly Canadian-born persons and representing local transmission, persons born outside Canada represented 25/322 (7.9%) of diagnoses. Amongst these individuals, the median time from immigration was 40 years (IQR: 25–49), and of the 24 people whose specific country of birth was known, only 10 (41.7%) emigrated from high-incidence TB countries (**Figure 9-3**). Seven of the 25 (28.0%) individuals reported risk factors similar to those seen in Canadian-born cluster members, including HIV and substance-use. In contrast, among large genotypic clusters consisting of predominantly persons born outside Canada from similar regions of the world—clusters that epidemiologic field work suggests largely arose from LTBI reactivation—the median time from

immigration was 12 years (IQR: 3–18), all 119 (100%) people emigrated from high-incidence TB countries, and none reported HIV or substance-use as risk factors.

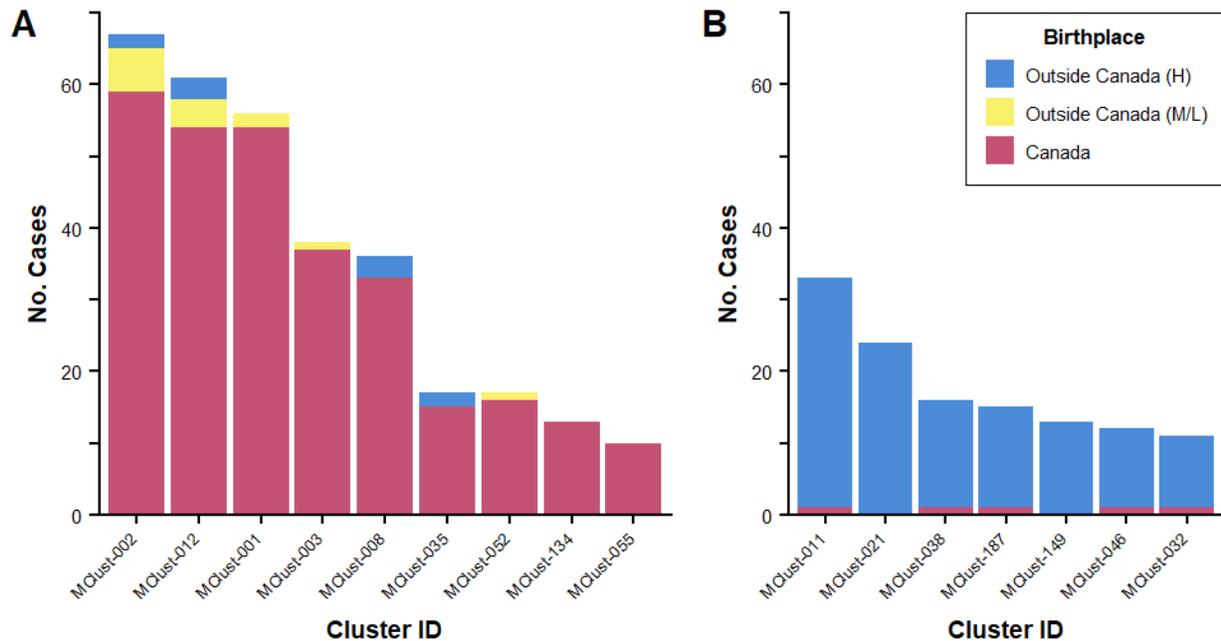


Figure 9-3. Number of tuberculosis cases for each large (≥ 10 persons) genotypic cluster. Large clusters in British Columbia, Canada (2005–2015) by cluster type: (A) predominantly Canadian-born and presumed to represent local transmission; (B) predominantly persons born outside Canada and presumed to largely represent reactivation of LTBI. Coloured to indicate persons born outside Canada (H—high-incidence countries, M/L—medium to low-incidence countries), or Canadian-born.

Discussion

Across Canada, 70% of new TB diagnoses occur in persons born outside Canada. These individuals come from all nine epidemiological regions defined by the World Health Organization, though each province has its own unique demographics.¹³ Between the years 2005 and 2012, there were 337,492 new Canadian permanent residents arriving in BC—one of the largest immigrant receiving provinces in Canada, with the majority (78.3%) coming from high TB incidence countries. The TB incidence rate in BC was 22.1 per 100,000 person-years in this population during this period.⁴¹⁶ The differences observed in absolute numbers and rate, as well

as time since immigration to active disease, and risk factors such as substance-use, underscore the notion that persons born outside Canada are not a homogeneous group when it comes to local transmission or risk factors.

Stratifying Individuals by Country of Birth Reveals Different Trends in TB Epidemiology

TB programs regularly report case counts and incidence rates in Canadian-born non-Indigenous, Canadian-born Indigenous (First Nations, Inuit, and Métis), and “foreign-born” populations. While this may be sufficient for high-level reporting and surveillance purposes, case investigations, program evaluations, and research should examine TB trends at a more granular level to obtain an accurate picture of TB epidemiology, including individual-level risk factors and estimates of recent, local transmission. As found in the study, splitting persons born outside of Canada into two groups—people born in high-incidence countries and people born in medium to low-incidence countries—yields a very different picture of TB epidemiology as it relates to foreign birth. Younger recent migrants from high-incidence countries largely represent LTBI reactivation, whereas long-time residents of Canada from lower incidence regions acquire TB locally. However, it should be noted that this study was limited to a single, albeit large and diverse Canadian province, and the results may not be generalizable to all low-incidence settings.

Recognizing the diversity within migrants and tailoring interventions to those at the highest risk for TB is clearly important; however, there is a fine balance between describing higher “risk” groups and developing targeted interventions while not reinforcing stigma. Stigma and concerns about immigration status amongst newcomers to Canada may delay individuals from seeking treatment.⁴¹⁰ Furthermore, the media’s portrayal of migrants as a potential public health risk contributes to this stigma.^{402,403} In reality, data from molecular epidemiology studies show that immigrants to low-incidence TB countries who are subsequently diagnosed with TB rarely transmit TB outside of their immediate family.^{404,405,417} Identification and quantification of risk factors beyond simply “foreign-born”, particularly in the context of TB molecular data will allow for a more clear understanding of TB transmission.

Conclusion

TB surveillance data and population research provides the evidence upon which many programs and policies are based.⁴¹⁸ Research examining public health interventions for populations have found that targeted strategies tailored to smaller, better defined subgroups are often more successful.⁴¹⁹ Change will require substantial evidence to define the optimal strategy to eliminate TB, but a first step towards this end is for TB surveillance programs and population studies to recognize the diversity amongst people with TB and to analyze data with the understanding that people born abroad represent a culturally, socially, and ethnically diverse population.

Disclaimer Statement

All inferences, opinions, and conclusions drawn in this study are those of the author, and do not reflect the opinions or policies of the Data Steward(s).

9.4 Knowledge translation

Studies suggest that it often takes many years before research findings are reflected in practice or translated into improved health outcomes for patients.⁴²⁰ To meet TB elimination goals, it is clear that we will need to reduce the time lag between research and implementation. During all stages of planning and conducting the research presented herein, knowledge translation and dissemination strategies were applied. The new knowledge generated from these projects is not only applicable to TB programs in BC, but also those across Canada and in other low-incidence settings. Accordingly, I have presented my research findings on 20 different occasions at local, provincial, and international meetings and conferences. Recently, results of this dissertation have been presented in a meeting with BC's Tuberculosis Strategic Plan Implementation Committee. Public-facing communication was achieved through the media, which included an online publication for World TB Day (2018) and a podcast interview. Full details can be seen in **Appendix B**. To date, three peer-reviewed original research articles have been published as a result of my work, with five more publications expected. I have also published a review and one commentary paper.

To translate key findings from the genotyping study to BC's TB program stakeholders, which required conversion of highly technical data into an intuitive and interpretable format, I prepared a genotyping summary report (**Appendix C**), the content of which I presented at BC's TB Clinical Leadership Meeting. This report has been disseminated to various laboratory and clinical stakeholders for use as training and reference material, and is now publicly available on the BCCDC website. Similar to the genotyping report, two documents based on the results of the genomic epidemiology studies are currently under preparation, one for BC and another for YT. Additionally, a training session for the use of genomic data to guide contact investigations will be held for staff at the Yukon Communicable Disease Control based on the recommendations presented in **Chapter 7**.

9.5 Future research and recommendations

The data obtained from genotyping and WGS of *Mtb* isolates in British Columbia (2005–2014) as part of this dissertation has provided significant insights to the population structure and transmission dynamics of *Mtb* in BC, providing an accurate baseline estimate of recent transmission rates, against which progress towards elimination can be measured. The genomic database created for these studies will form the basis for multiple future projects, and provide an important retrospective database against which prospective sequenced isolates can be compared.

9.5.1 Prospective provincial MIRU-VNTR genotyping

Prior to 2015, genotyping was only done on request from a clinician, generally in support of outbreak investigations and contact tracing efforts. The study results described in **Chapter 2** made the added value of universal genotyping clear, and as of 2015, TB genotyping has been part of routine laboratory operations. Future implementation-related work in this area includes: i) developing a means to automatically link case- and cluster-level information to genotyping results, allowing for epidemiologically-informed interpretation of the genotype data; ii) developing a reporting process that effectively communicates genotyping results in a timely manner to all stakeholders; and iii) incorporating genotypic data into BC's annual TB report.

9.5.2 Standardization of WGS bioinformatics pipelines

As WGS moves from the microbial research laboratory into routine clinical and public health practice, regulatory bodies will require standardized, clinically validated protocols for both WGS laboratory methods and interpretation of results. Currently, Public Health England is leading the way in laboratory accreditation of WGS for TB diagnosis and surveillance, and has developed and validated a bioinformatics pipeline for downstream analysis, which was used for the studies presented in this dissertation. However, many other research and public health laboratories have independently developed software pipelines and there is now a need for standardization of these computational approaches. Recently, a group of individuals involved world-wide in *Mtb* sequencing and analysis came together to jointly author a position paper on the minimum standards for a TB WGS bioinformatics pipeline—information necessary for both the reliable

interpretation of genome data and cross-laboratory comparison of data. Dr. Gardy and myself have been included in this group, and we are contributing our expertise in the form of recommendations around standardized reporting and visualization of results, both clinically and in the research community.

9.5.3 WGS as a tool for TB prevention

This work has provided a substantial database of over 1,500 *Mtb* genome sequences, all of which have been made publicly available through NCBI, as well as a list of insights into TB transmission that will be beneficial as BC plans for the implementation of prospective, real-time WGS. Yet, a number of challenges remain. First, laboratory results, including genomic data, and case-level information—key to interpreting the genomic data—are housed separately within databases that are not linked. Routine data linkage and ongoing curation of case- and isolate-level data requires significant personnel effort and the ability to link data across two discrete IT systems. Second, interpretation of WGS data in the context of epidemiological data is a manual process, requiring the time and expertise of a staff member able to understand both genomic data and the clinical and epidemiological complexities of TB. Ideally, an individual dedicated to this would work on TB genomic data, interpreting routine results and monitoring outbreaks and emerging strains of public health concern, such as drug-resistant strains. Lastly, communication of a novel type of complex laboratory data, such as genome sequence, is challenging, as many stakeholders are unfamiliar with genomic data and its interpretation. Work has begun to determine the most effective means of reporting WGS clustering data such that it is acceptable and interpretable to end-users and is able to support TB prevention programs and policies.

While the ten-year genomics studies presented herein have provided a broad overview of person-to-person spread of TB within the province, the next steps will be to examine transmission at a deeper level. Future studies will focus on developing models using epidemiologically-informed genomic data that will improve our ability to predict onward transmission at a case- and outbreak-level. Understanding the characteristics of individuals more likely to spread TB and being able to predict whether or not a cluster may go on to become a large outbreak are both

important future analyses that will support the development of evidence-based strategies for public health intervention.

9.6 Final Conclusions

The overall aim of this dissertation was to describe the molecular and genomic epidemiology of TB in British Columbia to improve our knowledge and understanding of tuberculosis transmission. While genotyping revealed considerable strain diversity—indicative of infections acquired outside BC—WGS provided a more accurate picture of TB dynamics in the province and provided unquestionable evidence of ongoing transmission within BC. Characterization of these transmission clusters provided new insights into sub-populations involved in person-to-person spread of TB, and the WGS data provided a more accurate estimate of LTBI reactivation in persons born outside Canada—key to obtaining funding for and efficiently delivering immigrant screening programs. In conclusion, the depth and breadth of the research presented herein, ranging from metropolitan centers to remote northern regions and including both adult and pediatric cases over a decade, means that the study outcomes are likely to be applicable to a variety of settings. Within BC the findings from this research have already begun to inform new policy and practice and are providing the basis for future studies.

8-9References

1. Styblo K. The relationship between the risk of tuberculosis infection and the risk of developing tuberculosis. *Bull Int Union Tuberc*. 1985;60:117–9.
2. Dye C. Tuberculosis 2000-2010: control, but not elimination. *Int J Tuberc Lung Dis*. 2000 Dec;4(12 Suppl 2):S146-152.
3. Ma M-J, Yang Y, Wang H-B, Zhu Y-F, Fang L-Q, An X-P, et al. Transmissibility of tuberculosis among school contacts: An outbreak investigation in a boarding middle school, China. *Infection, Genetics and Evolution*. 2015 Jun;32:148–55.
4. Mathema B, Andrews JR, Cohen T, Borgdorff MW, Behr M, Glynn JR, et al. Drivers of Tuberculosis Transmission. *J Infect Dis*. 2017 Nov 3;216(suppl_6):S644–53.
5. World Health Organization. *Global Tuberculosis Report 2017*. Geneva: World Health Organization; 2017.
6. Public Health Agency of Canada and Canadian Lung Association/Canadian Thoracic Society. *Canadian Tuberculosis Standards - 7th edition* [Internet]. 2014 [cited 2018 Oct 1]. Available from: <https://www.canada.ca/en/public-health/services/infectious-diseases/canadian-tuberculosis-standards-7th-edition/edition-22.html>
7. Public Health Agency of Canada Government of Canada. *Tuberculosis in Canada 2012 - Pre-Release – Public Health Agency of Canada* [Internet]. 2014 [cited 2015 Jun 15]. Available from: <http://www.phac-aspc.gc.ca/tbpc-latb/pubs/tbcan12pre/tab-eng.php#tab1>
8. BC Centre for Disease Control. *TB in British Columbia: Annual Surveillance Report 2014* [Internet]. 2016 [cited 2017 Jun 1]. Available from: http://www.bccdc.ca/resource-gallery/Documents/Statistics%20and%20Research/Statistics%20and%20Reports/TB/TB_Annual_Report_2014.pdf
9. BC Centre for Disease Control. *TB in British Columbia: Annual Surveillance Report 2012-2013* [Internet]. 2014 [cited 2017 Jul 7]. Available from: http://www.bccdc.ca/resource-gallery/Documents/Statistics%20and%20Research/Statistics%20and%20Reports/TB/TB_Annual_Report_2012-2013.pdf
10. BC Centre for Disease Control. *TB Annual Report 2005-2008*. Vancouver, Canada [Internet]. 2008 [cited 2017 Jun 1]. Available from: http://www.bccdc.ca/NR/rdonlyres/4E5E68CC-12CD-42A5-88DC-7CE57E31FEAE/0/2005_2008MultiAnnualTBReport_LowResAmended_Nov10.pdf
11. Menzies D, Lewis M, Oxlade O. Costs for tuberculosis care in Canada. *Can J Public Health*. 2008 Oct;99(5):391–6.

12. Health Canada and the Public Health Agency of Canada. Tuberculosis Prevention and Control in Canada - A Federal Framework for Action [Internet]. 2014 [cited 2015 May 25]. Available from: http://www.bccdc.ca/NR/rdonlyres/4E5E68CC-12CD-42A5-88DC-7CE57E31FEAE/0/2005_2008MultiAnnualTBReport_LowResAmended_Nov10.pdf
13. Vachon J, Gallant V, Siu W. Tuberculosis in Canada, 2016 [Internet]. Public Health Agency of Canada; 2018 Jan [cited 2018 Jul 2]. (Canada Communicable Disease Report Monthly). Report No.: Volume 44-3/4. Available from: https://www.canada.ca/en/public-health/services/reports-publications/canada-communicable-disease-report-ccdr/monthly-issue/2018-44/issue-3-4-march-1-2018/article-1-tuberculosis-2016.html?utm_source=lyris&utm_medium=email_en&utm_content=1&utm_campaign=ccdr_18-44-3-4
14. BC Communicable Disease Policy Advisory Committee. BC strategic plan for tuberculosis prevention, treatment, and control [Internet]. 2012 Jun [cited 2015 May 25]. Available from: http://www.bccdc.ca/NR/rdonlyres/371821DC-D135-4BC6-8AD9-4F09CF667B29/0/BC_Strategic_Plan_Tuberculosis.pdf
15. Sepkowitz KA. How contagious is tuberculosis? *Clin Infect Dis*. 1996 Nov;23(5):954–62.
16. Pieters J. Mycobacterium tuberculosis and the Macrophage: Maintaining a Balance. *Cell Host & Microbe*. 2008 Jun 12;3(6):399–407.
17. McClean CM, Tobin DM. Macrophage form, function, and phenotype in mycobacterial infection: lessons from tuberculosis and other diseases. *Pathog Dis* [Internet]. 2016 Oct 1 [cited 2018 Apr 15];74(7). Available from: <https://academic.oup.com/femspd/article/74/7/ftw068/2197856>
18. Knechel NA. Tuberculosis: Pathophysiology, Clinical Features, and Diagnosis. *Crit Care Nurse*. 2009 Apr 1;29(2):34–43.
19. Kaplan G, Post FA, Moreira AL, Wainwright H, Kreiswirth BN, Tanverdi M, et al. Mycobacterium tuberculosis Growth at the Cavity Surface: a Microenvironment with Failed Immunity. *Infect Immun*. 2003 Dec 1;71(12):7099–108.
20. Young DB, Gideon HP, Wilkinson RJ. Eliminating latent tuberculosis. *Trends in Microbiology*. 2009 May;17(5):183–8.
21. Kasprowicz VO, Churchyard G, Lawn SD, Squire SB, Lalvani A. Diagnosing Latent Tuberculosis in High-Risk Individuals: Rising to the Challenge in High-Burden Areas. *J Infect Dis*. 2011 Nov 15;204(Suppl 4):S1168–78.
22. Hartman-Adams H, Clark K, Juckett G. Update on latent tuberculosis infection. *Am Fam Physician*. 2014 Jun 1;89(11):889–96.

23. Trajman A, Steffen RE, Menzies D. Interferon-Gamma Release Assays versus Tuberculin Skin Testing for the Diagnosis of Latent Tuberculosis Infection: An Overview of the Evidence [Internet]. *Pulmonary Medicine*. 2013 [cited 2018 Apr 17]. Available from: <https://www.hindawi.com/journals/pm/2013/601737/>
24. Ayub A, Yale SH, Reed KD, Nasser RM, Gilbert SR. Testing for Latent Tuberculosis. *Clin Med Res*. 2004 Aug;2(3):191–4.
25. Cobelens FG, Egwaga SM, Ginkel T, Muwinge H, Matee MI, Borgdorff MW. Tuberculin Skin Testing in Patients with HIV Infection: Limited Benefit of Reduced Cutoff Values. *Clin Infect Dis*. 2006 Sep 1;43(5):634–9.
26. Farhat M, Greenaway C, Pai M, Menzies D. False-positive tuberculin skin tests: what is the absolute effect of BCG and non-tuberculous mycobacteria? [Review Article]. *The International Journal of Tuberculosis and Lung Disease*. 2006 Nov 1;10(11):1192–204.
27. Kunst H. Diagnosis of latent tuberculosis infection: The potential role of new technologies. *Respiratory Medicine*. 2006 Dec 1;100(12):2098–106.
28. Denkinger CM, Dheda K, Pai M. Guidelines on interferon- γ release assays for tuberculosis infection: concordance, discordance or confusion? *Clinical Microbiology and Infection*. 2011 Jun 1;17(6):806–14.
29. American Thoracic Society. Diagnostic Standards and Classification of Tuberculosis in Adults and Children. *Am J Respir Crit Care Med*. 2000 Apr 1;161(4):1376–95.
30. James BW, Williams A, Marsh PD. The physiology and pathogenicity of *Mycobacterium tuberculosis* grown under controlled conditions in a defined medium. *J Appl Microbiol*. 2000 Apr;88(4):669–77.
31. Lobue P, Menzies D. Treatment of latent tuberculosis infection: An update. *Respirology*. 2010 May 1;15(4):603–22.
32. Horsburgh CR, Goldberg S, Bethel J, Chen S, Colson PW, Hirsch-Moverman Y, et al. Latent TB Infection Treatment Acceptance and Completion in the United States and Canada. *CHEST*. 2010 Feb 1;137(2):401–9.
33. World Health Organization, Stop TB Initiative (World Health Organization), editors. *Treatment of tuberculosis: guidelines*. 4th ed. Geneva: World Health Organization; 2010. 147 p.
34. Faustini A, Hall AJ, Perucci CA. Risk factors for multidrug resistant tuberculosis in Europe: a systematic review. *Thorax*. 2006 Feb 1;61(2):158–63.
35. Hoagland DT, Liu J, Lee RB, Lee RE. New agents for the treatment of drug-resistant *Mycobacterium tuberculosis*. *Advanced Drug Delivery Reviews*. 2016 Jul 1;102:55–72.

36. Allix-Béguec C, Arandjelovic I, Bi L, Bonnett M, Bradley P, Cabibbe A, et al. Prediction of Susceptibility to First-Line Tuberculosis Drugs by DNA Sequencing. *N Engl J Med* (Accepted Aug 2018).
37. Brites D, Gagneux S. Co-evolution of *Mycobacterium tuberculosis* and *Homo sapiens*. *Immunological Reviews*. 264(1):6–24.
38. Vayr F, Martin-Blondel G, Savall F, Soulat J-M, Deffontaines G, Herin F. Occupational exposure to human *Mycobacterium bovis* infection: A systematic review. *PLOS Neglected Tropical Diseases*. 2018 Jan 16;12(1):e0006208.
39. Gagneux S. Host–pathogen coevolution in human tuberculosis. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*. 2012 Mar 19;367(1590):850–9.
40. Firdessa R, Berg S, Hailu E, Schelling E, Gumi B, Erenso G, et al. Mycobacterial Lineages Causing Pulmonary and Extrapulmonary Tuberculosis, Ethiopia. *Emerg Infect Dis*. 2013 Mar;19(3):460–3.
41. Gagneux S, DeRiemer K, Van T, Kato-Maeda M, Jong BC de, Narayanan S, et al. Variable host–pathogen compatibility in *Mycobacterium tuberculosis*. *PNAS*. 2006 Feb 21;103(8):2869–73.
42. Coscolla M, Gagneux S. Does *M. tuberculosis* genomic diversity explain disease diversity? *Drug Discov Today Dis Mech*. 2010;7(1):e43–59.
43. Hanekom M, Gey van Pittius NC, McEvoy C, Victor TC, Van Helden PD, Warren RM. *Mycobacterium tuberculosis* Beijing genotype: A template for success. *Tuberculosis*. 2011 Nov 1;91(6):510–23.
44. Coscolla M, Gagneux S. Consequences of genomic diversity in *Mycobacterium tuberculosis*. *Semin Immunol*. 2014 Dec;26(6):431–44.
45. Sarkar R, Lenders L, Wilkinson KA, Wilkinson RJ, Nicol MP. Modern Lineages of *Mycobacterium tuberculosis* Exhibit Lineage-Specific Patterns of Growth and Cytokine Induction in Human Monocyte-Derived Macrophages. *PLoS One* [Internet]. 2012 Aug 16 [cited 2018 Jul 22];7(8). Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3420893/>
46. Click ES, Moonan PK, Winston CA, Cowan LS, Oeltmann JE. Relationship Between *Mycobacterium tuberculosis* Phylogenetic Lineage and Clinical Site of Tuberculosis. *Clin Infect Dis*. 2012 Jan 15;54(2):211–9.
47. Ford CB, Shah RR, Maeda MK, Gagneux S, Murray MB, Cohen T, et al. *Mycobacterium tuberculosis* mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nature Genetics*. 2013 Jul;45(7):784–90.

48. Turner RD, Bothamley GH. Cough and the Transmission of Tuberculosis. *J Infect Dis.* 2015 May 1;211(9):1367–72.
49. Houk VN, Baker JH, Sorensen K, Kent DC. The epidemiology of tuberculosis infection in a closed environment. *Arch Environ Health.* 1968 Jan;16(1):26–35.
50. Rodrigo T, Caylà JA, García de Olalla P, Galdós-Tangüis H, Jansà JM, Miranda P, et al. Characteristics of tuberculosis patients who generate secondary cases. *Int J Tuberc Lung Dis.* 1997 Aug;1(4):352–7.
51. Hernández-Garduño E, Cook V, Kunimoto D, Elwood RK, Black WA, FitzGerald JM. Transmission of tuberculosis from smear negative patients: a molecular epidemiology study. *Thorax.* 2004 Apr;59(4):286–90.
52. Tostmann A, Kik SV, Kalisvaart NA, Sebek MM, Verver S, Boeree MJ, et al. Tuberculosis Transmission by Patients with Smear-Negative Pulmonary Tuberculosis in a Large Cohort in The Netherlands. *Clin Infect Dis.* 2008 Nov 1;47(9):1135–42.
53. Golub JE, Bur S, Cronin WA, Gange S, Baruch N, Comstock GW, et al. Delayed tuberculosis diagnosis and tuberculosis transmission. *Int J Tuberc Lung Dis.* 2006 Jan;10(1):24–30.
54. Cheng S, Chen W, Yang Y, Chu P, Liu X, Zhao M, et al. Effect of Diagnostic and Treatment Delay on the Risk of Tuberculosis Transmission in Shenzhen, China: An Observational Cohort Study, 1993–2010. *PLOS ONE.* 2013 Jun 27;8(6):e67516.
55. Yates TA, Tanser F, Abubakar I. Plan Beta for tuberculosis: it's time to think seriously about poorly ventilated congregate settings [Internet]. 2016 [cited 2018 May 16]. Available from: <http://www.ingentaconnect.com/content/iatld/ijtld/2016/00000020/00000001/art00005#>
56. Trunz BB, Fine P, Dye C. Effect of BCG vaccination on childhood tuberculous meningitis and miliary tuberculosis worldwide: a meta-analysis and assessment of cost-effectiveness. *The Lancet.* 2006 Apr;367(9517):1173–80.
57. McIvor A, Koornhof H, Kana BD. Relapse, re-infection and mixed infections in tuberculosis disease. *Pathogens and Disease* [Internet]. 2017 Apr 1 [cited 2018 Jul 12];75(3). Available from: <https://academic.oup.com/femspd/article-lookup/doi/10.1093/femspd/ftx020>
58. Whalen CC, Zalwango S, Chiunda A, Malone L, Eisenach K, Joloba M, et al. Secondary Attack Rate of Tuberculosis in Urban Households in Kampala, Uganda. *PLOS ONE.* 2011 Feb 14;6(2):e16137.

59. Morrison J, Pai M, Hopewell PC. Tuberculosis and latent tuberculosis infection in close contacts of people with pulmonary tuberculosis in low-income and middle-income countries: a systematic review and meta-analysis. *The Lancet Infectious Diseases*. 2008 Jun 1;8(6):359–68.
60. Bothamley GH. Smoking and tuberculosis: a chance or causal association? *Thorax*. 2005 Jul 1;60(7):527–8.
61. Deiss RG, Rodwell TC, Garfein RS. Tuberculosis and Illicit Drug Use: Review and Update. *Clinical Infectious Diseases*. 2009;48(1):72–82.
62. Rehm J, Samokhvalov AV, Neuman MG, Room R, Parry C, Lönnroth K, et al. The association between alcohol use, alcohol use disorders and tuberculosis (TB). A systematic review. *BMC Public Health*. 2009 Dec 5;9:450.
63. Cegielski JP, McMurray DN. The relationship between malnutrition and tuberculosis: evidence from studies in humans and experimental animals. *Int J Tuberc Lung Dis*. 2004 Mar;8(3):286–98.
64. Corbett EL, Steketee RW, ter Kuile FO, Latif AS, Kamali A, Hayes RJ. HIV-1/AIDS and the control of other infectious diseases in Africa. *The Lancet*. 2002 Jun;359(9324):2177–87.
65. Cobat A, Barrera LF, Henao H, Arbeláez P, Abel L, García LF, et al. Tuberculin Skin Test Reactivity Is Dependent on Host Genetic Background in Colombian Tuberculosis Household Contacts. *Clin Infect Dis*. 2012 Apr 1;54(7):968–71.
66. Murray JF. A century of tuberculosis. *Am J Respir Crit Care Med*. 2004 Jun 1;169(11):1181–6.
67. Fox GJ, Barry SE, Britton WJ, Marks GB. Contact investigation for tuberculosis: a systematic review and meta-analysis. *Eur Respir J*. 2013 Jan;41(1):140–56.
68. Cook VJ, Shah L, Gardy J, Bourgeois A-C. Recommendations on modern contact investigation methods for enhancing tuberculosis control. *Int J Tuberc Lung Dis*. 2012;16(3):297–305.
69. Bryant JM, Schürch AC, van Deutekom H, Harris SR, de Beer JL, de Jager V, et al. Inferring patient to patient transmission of *Mycobacterium tuberculosis* from whole genome sequencing data. *BMC Infect Dis*. 2013 Feb 27;13:110.
70. Pevzner ES, Robison S, Donovan J, Allis D, Spitters C, Friedman R, et al. Tuberculosis Transmission and Use of Methamphetamines in Snohomish County, WA, 1991–2006. *American Journal of Public Health*. 2010 Dec;100(12):2481–6.

71. Asghar RJ, Patlan DE, Miner MC, Rhodes HD, Solages A, Katz DJ, et al. Limited Utility of Name-Based Tuberculosis Contact Investigations among Persons Using Illicit Drugs: Results of an Outbreak Investigation. *J Urban Health*. 2009 Sep;86(5):776–80.
72. Outhred AC, Holmes N, Sadsad R, Martinez E, Jelfs P, Hill-Cawthorne GA, et al. Identifying Likely Transmission Pathways within a 10-Year Community Outbreak of Tuberculosis by High-Depth Whole Genome Sequencing. *PLoS ONE*. 2016;11(3):e0150550.
73. Walker TM, Lalor MK, Broda A, Ortega LS, Morgan M, Parker L, et al. Assessment of Mycobacterium tuberculosis transmission in Oxfordshire, UK, 2007–12, with whole pathogen genome sequences: an observational study. *The Lancet Respiratory Medicine*. 2014 Apr 1;2(4):285–92.
74. Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L, Brodtkin E, et al. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N Engl J Med*. 2011 Feb 24;364(8):730–9.
75. Mears J, Abubakar I, Crisp D, Maguire H, Innes JA, Lilley M, et al. Prospective evaluation of a complex public health intervention: lessons from an initial and follow-up cross-sectional survey of the tuberculosis strain typing service in England. *BMC Public Health*. 2014 Oct 2;14:1023.
76. Teeter LD, Kammerer JS, Ghosh S, Nguyen DTM, Vempaty P, Tapia J, et al. Evaluation of 24-locus MIRU-VNTR genotyping in Mycobacterium tuberculosis cluster investigations in four jurisdictions in the United States, 2006–2010. *Tuberculosis*. 2017 Sep 1;106:9–15.
77. Supply P, Allix C, Lesjean S, Cardoso-Oelemann M, Rüsch-Gerdes S, Willery E, et al. Proposal for Standardization of Optimized Mycobacterial Interspersed Repetitive Unit-Variable-Number Tandem Repeat Typing of Mycobacterium tuberculosis. *J Clin Microbiol*. 2006 Dec 1;44(12):4498–510.
78. Gallant V, Duvvuri V, McGuire M. Tuberculosis in Canada - Summary 2015. *Can Commun Dis Rep*. 2017 Mar 2;43(3):77–82.
79. World Health Organization. Tuberculosis country profiles [Internet]. [cited 2018 Jul 8]. Available from: <http://www.who.int/tb/country/data/profiles/en/>
80. Public Health Ontario. Tuberculosis: Ontario Provincial Report, 2012 [Internet]. [cited 2018 Jul 13]. Available from: https://www.publichealthontario.ca/en/eRepository/Tuberculosis_Ontario_Provincial_Report_2012.pdf

81. Rivest P, Agence de la santé et des services sociaux de Montréal (Québec), Secteur Infection et intoxication dans la communauté, Bibliothèque numérique canadienne (Firme). Épidémiologie de la tuberculose au Québec de 2008 à 2011 [Internet]. 2014 [cited 2018 Jul 14]. Available from: <http://www.deslibris.ca/ID/243419>
82. Bourgeois A-C, Zulz T, Bruce MG, Stenz F, Koch A, Parkinson A, et al. Tuberculosis in the Circumpolar Region, 2006–2012. *The International Journal of Tuberculosis and Lung Disease*. 2018 Jun 1;22(6):641–8.
83. Public Health Agency of Canada. Tuberculosis in Canada 2012 [Internet]. Ottawa (ON); 2014. Available from: <http://www.phac-aspc.gc.ca/tbpc-latb/pubs/tbcan12/index-eng.php>
84. Public Health Agency of Canada. Tuberculosis in Canada 2014 – Pre-release [Internet]. 2016 [cited 2018 Jun 21]. Available from: <http://healthycanadians.gc.ca/publications/diseases-conditions-maladies-affections/tuberculosis-2014-tuberculose/index-eng.php#a4d>
85. Zignol M, Dean AS, Falzon D, van Gemert W, Wright A, van Deun A, et al. Twenty Years of Global Surveillance of Antituberculosis-Drug Resistance. *New England Journal of Medicine*. 2016 Sep 15;375(11):1081–9.
86. Moonan PK, Ghosh S, Oeltmann JE, Kammerer JS, Cowan LS, Navin TR. Using Genotyping and Geospatial Scanning to Estimate Recent Mycobacterium tuberculosis Transmission, United States. *Emerg Infect Dis*. 2012 Mar;18(3):458–65.
87. Yeo IKT, Tannenbaum T, Scott AN, Kozak R, Behr MA, Thibert L, et al. Contact Investigation and Genotyping to Identify Tuberculosis Transmission to Children: The Pediatric Infectious Disease Journal. 2006 Nov;25(11):1037–43.
88. Small PM, Hopewell PC, Singh SP, Paz A, Parsonnet J, Ruston DC, et al. The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods. *N Engl J Med*. 1994 Jun 16;330(24):1703–9.
89. Alland D, Kalkut GE, Moss AR, McAdam RA, Hahn JA, Bosworth W, et al. Transmission of Tuberculosis in New York City -- An Analysis by DNA Fingerprinting and Conventional Epidemiologic Methods. *New England Journal of Medicine*. 1994 Jun 16;330(24):1710–6.
90. Jasmer RM, Roemer M, Hamilton J, Bunter J, Braden CR, Shinnick TM, et al. A Prospective, Multicenter Study of Laboratory Cross-Contamination of Mycobacterium tuberculosis Cultures. *Emerg Infect Dis*. 2002 Nov;8(11):1260–3.
91. Bifani PJ, Mathema B, Liu Z, Moghazeh SL, Shopsin B, Tempalski B, et al. Identification of a W Variant Outbreak of Mycobacterium tuberculosis via Population-Based Molecular Epidemiology. *JAMA*. 1999 Dec 22;282(24):2321–7.

92. Pepperell CS, Granka JM, Alexander DC, Behr MA, Chui L, Gordon J, et al. Dispersal of *Mycobacterium tuberculosis* via the Canadian fur trade. *Proceedings of the National Academy of Sciences*. 2011;108(16):6526–6531.
93. Interrante JD, Haddad MB, Kim L, Gandhi NR. Exogenous Reinfection as a Cause of Late Recurrent Tuberculosis in the United States. *Annals ATS*. 2015 Sep 1;12(11):1619–26.
94. Olive DM, Bean P. Principles and applications of methods for DNA-based typing of microbial organisms. *J Clin Microbiol*. 1999 Jun;37(6):1661–9.
95. Thierry D, Brisson-Noël A, Vincent-Lévy-Frébault V, Nguyen S, Guesdon JL, Gicquel B. Characterization of a *Mycobacterium tuberculosis* insertion sequence, IS6110, and its application in diagnosis. *J Clin Microbiol*. 1990 Dec;28(12):2668–73.
96. Lanotte P. Molecular Epidemiology of Tuberculosis. In: Morand S, Beaudreau F, Cabaret J, editors. *New Frontiers of Molecular Epidemiology of Infectious Diseases* [Internet]. Dordrecht: Springer Netherlands; 2012 [cited 2018 Mar 24]. p. 125–47. Available from: http://www.springerlink.com/index/10.1007/978-94-007-2114-2_7
97. Roychowdhury T, Mandal S, Bhattacharya A. Analysis of IS6110 insertion sites provide a glimpse into genome evolution of *Mycobacterium tuberculosis*. *Scientific Reports* [Internet]. 2015 Dec [cited 2018 Mar 24];5(1). Available from: <http://www.nature.com/articles/srep12567>
98. de Boer AS, Borgdorff MW, de Haas PE, Nagelkerke NJ, van Embden JD, van Soolingen D. Analysis of rate of change of IS6110 RFLP patterns of *Mycobacterium tuberculosis* based on serial patient isolates. *J Infect Dis*. 1999 Oct;180(4):1238–44.
99. Warren RM, van der Spuy GD, Richardson M, Beyers N, Booysen C, Behr MA, et al. Evolution of the IS6110-based restriction fragment length polymorphism pattern during the transmission of *Mycobacterium tuberculosis*. *J Clin Microbiol*. 2002 Apr;40(4):1277–82.
100. Warren RM, van der Spuy GD, Richardson M, Beyers N, Borgdorff MW, Behr MA, et al. Calculation of the stability of the IS6110 banding pattern in patients with persistent *Mycobacterium tuberculosis* disease. *J Clin Microbiol*. 2002 May;40(5):1705–8.
101. Kato-Maeda M, Metcalfe JZ, Flores L. Genotyping of *Mycobacterium tuberculosis*: application in epidemiologic studies. *Future Microbiol*. 2011 Feb;6(2):203–16.
102. Cowan LS, Mosher L, Diem L, Massey JP, Crawford JT. Variable-Number Tandem Repeat Typing of *Mycobacterium tuberculosis* Isolates with Low Copy Numbers of IS6110 by Using *Mycobacterial Interspersed Repetitive Units*. *Journal of Clinical Microbiology*. 2002 May 1;40(5):1592–602.

103. Supply P, Mazars E, Lesjean S, Vincent V, Gicquel B, Locht C. Variable human minisatellite-like regions in the Mycobacterium tuberculosis genome. *Molecular Microbiology*. 2000 May 1;36(3):762–71.
104. Supply P, Magdalena J, Himpens S, Locht C. Identification of novel intergenic repetitive units in a mycobacterial two-component system operon. *Molecular Microbiology*. 1997 Dec;26(5):991–1003.
105. de Beer JL, Kremer K, Ködmön C, Supply P, Soolingen D van, Tuberculosis 2009 the GN for the MS of. First Worldwide Proficiency Study on Variable-Number Tandem-Repeat Typing of Mycobacterium tuberculosis Complex Strains. *J Clin Microbiol*. 2012 Mar 1;50(3):662–9.
106. Applied Biosystems. GeneMapper® Software v4.1 - Product Bulletin [Internet]. [cited 2018 Mar 30]. Available from: http://www3.appliedbiosystems.com/cms/groups/mcb_marketing/documents/generaldocuments/cms_077415.pdf
107. Mycobacterial Interspersed Repetitive Units (MIRU) typing [Internet]. 2012 [cited 2018 Mar 30]. Available from: <http://www.applied-maths.com/applications/mycobacterial-interspersed-repetitive-units-miru-typing>
108. de Beer JL, Akkerman OW, Schürch AC, Mulder A, Werf TS van der, Zanden AGM van der, et al. Optimization of Standard In-House 24-Locus Variable-Number Tandem-Repeat Typing for Mycobacterium tuberculosis and Its Direct Application to Clinical Material. *J Clin Microbiol*. 2014 May 1;52(5):1338–42.
109. Zanden AGM van der, Kremer K, Schouls LM, Caimi K, Cataldi A, Hulleman A, et al. Improvement of Differentiation and Interpretability of Spoligotyping for Mycobacterium tuberculosis Complex Isolates by Introduction of New Spacer Oligonucleotides. *J Clin Microbiol*. 2002 Dec 1;40(12):4628–39.
110. Cowan LS, Diem L, Brake MC, Crawford JT. Transfer of a Mycobacterium tuberculosis genotyping method, Spoligotyping, from a reverse line-blot hybridization, membrane-based assay to the Luminex multianalyte profiling system. *J Clin Microbiol*. 2004 Jan;42(1):474–7.
111. Dale JW, Brittain D, Cataldi AA, Cousins D, Crawford JT, Driscoll J, et al. Spacer oligonucleotide typing of bacteria of the Mycobacterium tuberculosis complex: recommendations for standardised nomenclature [The Language of Our Science]. *The International Journal of Tuberculosis and Lung Disease*. 2001 Mar 1;5(3):216–9.
112. Allix-Béguec C, Fauville-Dufaux M, Supply P. Three-Year Population-Based Evaluation of Standardized Mycobacterial Interspersed Repetitive-Unit-Variable-Number Tandem-Repeat Typing of Mycobacterium tuberculosis. *J Clin Microbiol*. 2008 Apr 1;46(4):1398–406.

113. Gori A, Bandera A, Marchetti G, Esposti AD, Catozzi L, Nardi GP, et al. Spoligotyping and Mycobacterium tuberculosis. *Emerg Infect Dis*. 2005 Aug;11(8):1242–8.
114. Stucki D, Brites D, Jeljeli L, Coscolla M, Liu Q, Trauner A, et al. Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages. *Nat Genet*. 2016 Oct 31;
115. Lagos J, Couvin D, Arata L, Tognarelli J, Aguayo C, Leiva T, et al. Analysis of Mycobacterium tuberculosis Genotypic Lineage Distribution in Chile and Neighboring Countries. *PLOS ONE*. 2016 Aug 12;11(8):e0160434.
116. Santos ACB, Gaspareto RM, Viana BHJ, Mendes NH, Pandolfi JRC, Cardoso RF, et al. Mycobacterium tuberculosis population structure shift in a 5-year molecular epidemiology surveillance follow-up study in a low endemic agro-industrial setting in São Paulo, Brazil. *Int J Mycobacteriol*. 2013 Sep;2(3):156–65.
117. Mehaffy C, Guthrie JL, Alexander DC, Stuart R, Rea E, Jamieson FB. Marked microevolution of a unique Mycobacterium tuberculosis strain in 17 years of ongoing transmission in a high risk population. *PLoS ONE*. 2014;9(11):e112928.
118. Black AT, Hamblion EL, Buttivant H, Anderson SR, Stone M, Casali N, et al. Tracking and responding to an outbreak of tuberculosis using MIRU-VNTR genotyping and whole genome sequencing as epidemiological tools. *J Public Health (Oxf)*. 2017 Jul 5;1–8.
119. Bryant JM, Harris SR, Parkhill J, Dawson R, Diacon AH, van Helden P, et al. Whole-genome sequencing to establish relapse or re-infection with Mycobacterium tuberculosis: a retrospective observational study. *Lancet Respir Med*. 2013 Dec;1(10):786–92.
120. Guerra-Assunção JA, Houben RMGJ, Crampin AC, Mzembe T, Mallard K, Coll F, et al. 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*. 2015 Apr 1;211(7):1154–63.
121. Hamblion EL, Menach AL, Anderson LF, Lalor MK, Brown T, Abubakar I, et al. Recent TB transmission, clustering and predictors of large clusters in London, 2010–2012: results from first 3 years of universal MIRU-VNTR strain typing. *Thorax*. 2016 Aug 1;71(8):749–56.
122. Gurjav U, Outhred AC, Jelfs P, McCallum N, Wang Q, Hill-Cawthorne GA, et al. Whole Genome Sequencing Demonstrates Limited Transmission within Identified Mycobacterium tuberculosis Clusters in New South Wales, Australia. *PLoS One* [Internet]. 2016 Oct 13 [cited 2018 Jul 22];11(10). Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5063377/>

123. Ferdinand S, Millet J, Accipe A, Cassadou S, Chaud P, Levy M, et al. Use of genotyping based clustering to quantify recent tuberculosis transmission in Guadeloupe during a seven years period: analysis of risk factors and access to health care. *BMC Infectious Diseases* [Internet]. 2013 Dec [cited 2018 Jul 22];13(1). Available from: <http://bmcinfectdis.biomedcentral.com/articles/10.1186/1471-2334-13-364>
124. Shea KM, Kammerer JS, Winston CA, Navin TR, Horsburgh. Estimated Rate of Reactivation of Latent Tuberculosis Infection in the United States, Overall and by Population Subgroup. *Am J Epidemiol*. 2014 Jan 15;179(2):216–25.
125. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, et al. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*. 1998 Jun 11;393(6685):537–44.
126. Fleischmann RD, Alland D, Eisen JA, Carpenter L, White O, Peterson J, et al. Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. *J Bacteriol*. 2002 Oct;184(19):5479–90.
127. National Center for Biotechnology Information. *M. tuberculosis* genomes [Internet]. [cited 2018 Jan 14]. Available from: <https://www.ncbi.nlm.nih.gov/genome/genomes/166>
128. National Center for Biotechnology Information. Sequence Read Archive [Internet]. [cited 2018 Jan 14]. Available from: <https://www.ncbi.nlm.nih.gov/sra>
129. World Health Organization. Global Tuberculosis Report 2015 [Internet]. 2015 [cited 2017 Jun 1]. Available from: http://apps.who.int/iris/bitstream/10665/191102/1/9789241565059_eng.pdf?ua=1
130. Corbett EL, Watt CJ, Walker N, Maher D, Williams BG, Raviglione MC, et al. The growing burden of tuberculosis: global trends and interactions with the HIV epidemic. *Arch Intern Med*. 2003 May 12;163(9):1009–21.
131. Mostowy S, Behr MA. The origin and evolution of *Mycobacterium tuberculosis*. *Clin Chest Med*. 2005 Jun;26(2):207–16, v–vi.
132. Veyrier F, Pletzer D, Turenne C, Behr MA. Phylogenetic detection of horizontal gene transfer during the step-wise genesis of *Mycobacterium tuberculosis*. *BMC Evol Biol*. 2009 Aug 10;9:196.
133. Pepperell CS, Casto AM, Kitchen A, Granka JM, Cornejo OE, Holmes EC, et al. The role of selection in shaping diversity of natural *M. tuberculosis* populations. *PLoS Pathog*. 2013 Aug;9(8):e1003543.
134. Eldholm V, Balloux F. Antimicrobial Resistance in *Mycobacterium tuberculosis*: The Odd One Out. *Trends Microbiol*. 2016 Aug;24(8):637–48.

135. Lönnroth K, Migliori GB, Abubakar I, D'Ambrosio L, de Vries G, Diel R, et al. Towards tuberculosis elimination: an action framework for low-incidence countries. *Eur Respir J*. 2015 Apr;45(4):928–52.
136. Iseman MD, Bentz RR, Fraser RI, Locks MO, Ostrow JH, Sewell EM. Guidelines for the investigation and management of tuberculosis contacts. *Am Rev Respir Dis*. 1976 Aug;114(2):459–63.
137. National Tuberculosis Controllers Association, Centers for Disease Control and Prevention (CDC). Guidelines for the investigation of contacts of persons with infectious tuberculosis. Recommendations from the National Tuberculosis Controllers Association and CDC. *MMWR Recomm Rep*. 2005 Dec 16;54(RR-15):1–47.
138. McNabb SJN, Kammerer JS, Hickey AC, Braden CR, Shang N, Rosenblum LS, et al. Added Epidemiologic Value to Tuberculosis Prevention and Control of the Investigation of Clustered Genotypes of Mycobacterium tuberculosis Isolates. *Am J Epidemiol*. 2004 Sep 15;160(6):589–97.
139. Nikolayevskyy V, Kranzer K, Niemann S, Drobniowski F. Whole genome sequencing of Mycobacterium tuberculosis for detection of recent transmission and tracing outbreaks: A systematic review. *Tuberculosis (Edinb)*. 2016 May;98:77–85.
140. Jamieson FB, Teatero S, Guthrie JL, Neemuchwala A, Fittipaldi N, Mehaffy C. 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*. 2014 Oct;52(10):3795–8.
141. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. *Genomics*. 2016 Jan;107(1):1–8.
142. Parkhill J, Wren BW. Bacterial epidemiology and biology--lessons from genome sequencing. *Genome Biol*. 2011 Oct 24;12(10):230.
143. Croucher NJ, Didelot X. The application of genomics to tracing bacterial pathogen transmission. *Curr Opin Microbiol*. 2015 Feb;23:62–7.
144. Didelot X, Gardy J, Colijn C. Bayesian Inference of Infectious Disease Transmission from Whole-Genome Sequence Data. *Mol Biol Evol*. 2014 Jul 1;31(7):1869–79.
145. Hall M, Woolhouse M, Rambaut A. Epidemic Reconstruction in a Phylogenetics Framework: Transmission Trees as Partitions of the Node Set. *PLoS Comput Biol*. 2015 Dec;11(12):e1004613.
146. Worby CJ, O'Neill PD, Kypraios T, Robotham JV, De Angelis D, Cartwright EJP, et al. Reconstructing transmission trees for communicable diseases using densely sampled genetic data. *Ann Appl Stat*. 2016 Mar;10(1):395–417.

147. Schürch AC, Kremer K, Kiers A, Daviena O, Boeree MJ, Siezen RJ, et al. The tempo and mode of molecular evolution of *Mycobacterium tuberculosis* at patient-to-patient scale. *Infect Genet Evol.* 2010 Jan;10(1):108–14.
148. Cook VJ, Sun SJ, Tapia J, Muth SQ, Argüello DF, Lewis BL, et al. Transmission Network Analysis in Tuberculosis Contact Investigations. *J Infect Dis.* 2007 Nov 15;196(10):1517–27.
149. Walker TM, Ip CLC, Harrell RH, Evans JT, Kapatai G, Dediccoat MJ, et al. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *Lancet Infect Dis.* 2013 Feb;13(2):137–46.
150. Duchêne S, Holt KE, Weill F-X, Le Hello S, Hawkey J, Edwards DJ, et al. Genome-scale rates of evolutionary change in bacteria. *Microb Genom.* 2016 Nov;2(11):e000094.
151. Gilchrist CA, Turner SD, Riley MF, Petri WA, Hewlett EL. Whole-genome sequencing in outbreak analysis. *Clin Microbiol Rev.* 2015 Jul;28(3):541–63.
152. Sintchenko V, Holmes EC. The role of pathogen genomics in assessing disease transmission. *BMJ.* 2015 May 11;350:h1314.
153. Bentley SD, Parkhill J. Genomic perspectives on the evolution and spread of bacterial pathogens. *Proc Biol Sci.* 2015 Dec 22;282(1821):20150488.
154. Deng X, den Bakker HC, Hendriksen RS. Genomic Epidemiology: Whole-Genome-Sequencing-Powered Surveillance and Outbreak Investigation of Foodborne Bacterial Pathogens. *Annu Rev Food Sci Technol.* 2016;7:353–74.
155. Schürch AC, Kremer K, Daviena O, Kiers A, Boeree MJ, Siezen RJ, et al. High-resolution typing by integration of genome sequencing data in a large tuberculosis cluster. *J Clin Microbiol.* 2010 Sep;48(9):3403–6.
156. Clark TG, Mallard K, Coll F, Preston M, Assefa S, Harris D, et al. Elucidating emergence and transmission of multidrug-resistant tuberculosis in treatment experienced patients by whole genome sequencing. *PLoS ONE.* 2013;8(12):e83012.
157. Kato-Maeda M, Ho C, Passarelli B, Banaei N, Grinsdale J, Flores L, et al. Use of whole genome sequencing to determine the microevolution of *Mycobacterium tuberculosis* during an outbreak. *PLoS ONE.* 2013;8(3):e58235.
158. Roetzer A, Diel R, Kohl TA, Rückert C, Nübel U, Blom J, et al. Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiological study. *PLoS Med.* 2013;10(2):e1001387.

159. Török ME, Reuter S, Bryant J, Köser CU, Stinchcombe SV, Nazareth B, et al. Rapid whole-genome sequencing for investigation of a suspected tuberculosis outbreak. *J Clin Microbiol.* 2013 Feb;51(2):611–4.
160. Kohl TA, Diel R, Harmsen D, Rothgänger J, Walter KM, Merker M, et al. Whole-genome-based *Mycobacterium tuberculosis* surveillance: a standardized, portable, and expandable approach. *J Clin Microbiol.* 2014 Jul;52(7):2479–86.
161. Luo T, Yang C, Peng Y, Lu L, Sun G, Wu J, et al. Whole-genome sequencing to detect recent transmission of *Mycobacterium tuberculosis* in settings with a high burden of tuberculosis. *Tuberculosis (Edinb).* 2014 Jul;94(4):434–40.
162. Pérez-Lago L, Comas I, Navarro Y, González-Candelas F, Herranz M, Bouza E, et al. Whole genome sequencing analysis of intrapatient microevolution in *Mycobacterium tuberculosis*: potential impact on the inference of tuberculosis transmission. *J Infect Dis.* 2014 Jan 1;209(1):98–108.
163. Coscolla M, Barry PM, Oeltmann JE, Koshinsky H, Shaw T, Cilnis M, et al. Genomic epidemiology of multidrug-resistant *Mycobacterium tuberculosis* during transcontinental spread. *J Infect Dis.* 2015 Jul 15;212(2):302–10.
164. Glynn JR, Guerra-Assunção JA, Houben RMGJ, Sichali L, Mzembe T, Mwaungulu LK, et al. Whole Genome Sequencing Shows a Low Proportion of Tuberculosis Disease Is Attributable to Known Close Contacts in Rural Malawi. *PLoS ONE.* 2015;10(7):e0132840.
165. Guerra-Assunção JA, Crampin AC, Houben RMGJ, Mzembe T, Mallard K, Coll F, et al. Large-scale whole genome sequencing of *M. tuberculosis* provides insights into transmission in a high prevalence area. *Elife.* 2015 Mar 3;4.
166. Lee RS, Radomski N, Proulx J-F, Manry J, McIntosh F, Desjardins F, et al. Reemergence and Amplification of Tuberculosis in the Canadian Arctic. *J Infect Dis.* 2015 Jun 15;211(12):1905–14.
167. Lee RS, Radomski N, Proulx J-F, Levade I, Shapiro BJ, McIntosh F, et al. Population genomics of *Mycobacterium tuberculosis* in the Inuit. *Proc Natl Acad Sci USA.* 2015 Nov 3;112(44):13609–14.
168. Regmi SM, Chaiprasert A, Kulawonganuchai S, Tongsimma S, Coker OO, Prammananan T, et al. Whole genome sequence analysis of multidrug-resistant *Mycobacterium tuberculosis* Beijing isolates from an outbreak in Thailand. *Mol Genet Genomics.* 2015 Oct;290(5):1933–41.

169. Stucki D, Ballif M, Bodmer T, Coscolla M, Maurer A-M, Droz S, et al. Tracking a tuberculosis outbreak over 21 years: strain-specific single-nucleotide polymorphism typing combined with targeted whole-genome sequencing. *J Infect Dis*. 2015 Apr 15;211(8):1306–16.
170. Witney AA, Gould KA, Arnold A, Coleman D, Delgado R, Dhillon J, et al. Clinical application of whole-genome sequencing to inform treatment for multidrug-resistant tuberculosis cases. *J Clin Microbiol*. 2015 May;53(5):1473–83.
171. Arnold A, Witney AA, Vergnano S, Roche A, Cosgrove CA, Houston A, et al. XDR-TB transmission in London: Case management and contact tracing investigation assisted by early whole genome sequencing. *J Infect*. 2016 Sep;73(3):210–8.
172. Land M, Hauser L, Jun S-R, Nookaew I, Leuze MR, Ahn T-H, et al. Insights from 20 years of bacterial genome sequencing. *Funct Integr Genomics*. 2015 Mar;15(2):141–61.
173. Cheng JM, Hiscoe L, Pollock SL, Hasselback P, Gardy JL, Parker R. A clonal outbreak of tuberculosis in a homeless population in the interior of British Columbia, Canada, 2008–2015. *Epidemiology & Infection*. 2015 Nov;143(15):3220–6.
174. Hatherell H-A, Didelot X, Pollock SL, Tang P, Crisan A, Johnston JC, et al. Declaring a tuberculosis outbreak over with genomic epidemiology. *Microbial Genomics* [Internet]. 2016 [cited 2016 Sep 19];2(5). Available from: <http://mgen.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000060>
175. Gardy JL, Naus M, Amlani A, Chung W, Kim H, Tan M, et al. Whole-Genome Sequencing of Measles Virus Genotypes H1 and D8 During Outbreaks of Infection Following the 2010 Olympic Winter Games Reveals Viral Transmission Routes. *J Infect Dis*. 2015 Nov 15;212(10):1574–8.
176. Fournier P-E, Drancourt M, Colson P, Rolain J-M, La Scola B, Raoult D. Modern clinical microbiology: new challenges and solutions. *Nat Rev Microbiol*. 2013 Aug;11(8):574–85.
177. Parkinson NJ, Maslau S, Ferneyhough B, Zhang G, Gregory L, Buck D, et al. Preparation of high-quality next-generation sequencing libraries from picogram quantities of target DNA. *Genome Res*. 2012 Jan;22(1):125–33.
178. Graham RMA, Doyle CJ, Jennison AV. Epidemiological typing of *Neisseria gonorrhoeae* and detection of markers associated with antimicrobial resistance directly from urine samples using next generation sequencing. *Sex Transm Infect*. 2017 Feb;93(1):65–7.
179. Andersson P, Klein M, Lilliebridge RA, Giffard PM. Sequences of multiple bacterial genomes and a *Chlamydia trachomatis* genotype from direct sequencing of DNA derived from a vaginal swab diagnostic specimen. *Clin Microbiol Infect*. 2013 Sep;19(9):E405–408.

180. Brown AC, Bryant JM, Einer-Jensen K, Holdstock J, Houniet DT, Chan JZM, et al. Rapid Whole-Genome Sequencing of Mycobacterium tuberculosis Isolates Directly from Clinical Samples. *J Clin Microbiol*. 2015 Jul;53(7):2230–7.
181. Loman NJ, Constantinidou C, Christner M, Rohde H, Chan JZ-M, Quick J, et al. A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxicogenic Escherichia coli O104:H4. *JAMA*. 2013 Apr 10;309(14):1502–10.
182. Doughty EL, Sergeant MJ, Adetifa I, Antonio M, Pallen MJ. Culture-independent detection and characterisation of Mycobacterium tuberculosis and M. africanum in sputum samples using shotgun metagenomics on a benchtop sequencer. *PeerJ*. 2014;2:e585.
183. Bell RL, Jarvis KG, Ottesen AR, McFarland MA, Brown EW. Recent and emerging innovations in Salmonella detection: a food and environmental perspective. *Microb Biotechnol*. 2016 May;9(3):279–92.
184. Pallen MJ. Diagnostic metagenomics: potential applications to bacterial, viral and parasitic infections. *Parasitology*. 2014 Dec;141(14):1856–62.
185. National Human Genome Research Institute. DNA Sequencing Costs: Data [Internet]. [cited 2018 Jan 14]. Available from: <https://www.genome.gov/27541954/DNA-Sequencing-Costs-Data>
186. McGrath M, Gey van Pittius NC, van Helden PD, Warren RM, Warner DF. Mutation rate and the emergence of drug resistance in Mycobacterium tuberculosis. *J Antimicrob Chemother*. 2014 Feb;69(2):292–302.
187. Hedge J, Wilson DJ. Bacterial phylogenetic reconstruction from whole genomes is robust to recombination but demographic inference is not. *MBio*. 2014 Nov 25;5(6):e02158.
188. Conlan S, Thomas PJ, Deming C, Park M, Lau AF, Dekker JP, et al. Single-molecule sequencing to track plasmid diversity of hospital-associated carbapenemase-producing Enterobacteriaceae. *Sci Transl Med*. 2014 Sep 17;6(254):254ra126.
189. Hardiman CA, Weingarten RA, Conlan S, Khil P, Dekker JP, Mathers AJ, et al. Horizontal Transfer of Carbapenemase-Encoding Plasmids and Comparison with Hospital Epidemiology Data. *Antimicrob Agents Chemother*. 2016 Aug;60(8):4910–9.
190. Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O, Villa L, et al. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob Agents Chemother*. 2014 Jul;58(7):3895–903.
191. Antipov D, Hartwick N, Shen M, Raiko M, Lapidus A, Pevzner PA. plasmidSPAdes: assembling plasmids from whole genome sequencing data. *Bioinformatics*. 2016 Nov 15;32(22):3380–7.

192. Croucher NJ, Harris SR, Grad YH, Hanage WP. Bacterial genomes in epidemiology--present and future. *Philos Trans R Soc Lond, B, Biol Sci.* 2013 Mar 19;368(1614):20120202.
193. Bryant J, Chewapreecha C, Bentley SD. Developing insights into the mechanisms of evolution of bacterial pathogens from whole-genome sequences. *Future Microbiol.* 2012 Nov;7(11):1283–96.
194. Köser CU, Holden MTG, Ellington MJ, Cartwright EJP, Brown NM, Ogilvy-Stuart AL, et al. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *N Engl J Med.* 2012 Jun 14;366(24):2267–75.
195. Phelan JE, Coll F, Bergval I, Anthony RM, Warren R, Sampson SL, et al. Recombination in *pe/ppe* genes contributes to genetic variation in *Mycobacterium tuberculosis* lineages. *BMC Genomics.* 2016 Feb 29;17:151.
196. Sekizuka T, Yamashita A, Murase Y, Iwamoto T, Mitarai S, Kato S, et al. TGS-TB: Total Genotyping Solution for *Mycobacterium tuberculosis* Using Short-Read Whole-Genome Sequencing. *PLoS ONE.* 2015;10(11):e0142951.
197. WGSANet. Whole Genome Sequence Analysis [Internet]. Available from: <http://www.wgsa.net>
198. Eppinger M, Pearson T, Koenig SSK, Pearson O, Hicks N, Agrawal S, et al. Genomic epidemiology of the Haitian cholera outbreak: a single introduction followed by rapid, extensive, and continued spread characterized the onset of the epidemic. *MBio.* 2014 Nov 4;5(6):e01721.
199. Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science.* 2014 Sep 12;345(6202):1369–72.
200. Aanensen DM, Feil EJ, Holden MTG, Dordel J, Yeats CA, Fedosejev A, et al. Whole-Genome Sequencing for Routine Pathogen Surveillance in Public Health: a Population Snapshot of Invasive *Staphylococcus aureus* in Europe. *MBio.* 2016 05;7(3).
201. Snitkin ES, Zelazny AM, Thomas PJ, Stock F, NISC Comparative Sequencing Program Group, Henderson DK, et al. Tracking a hospital outbreak of carbapenem-resistant *Klebsiella pneumoniae* with whole-genome sequencing. *Sci Transl Med.* 2012 Aug 22;4(148):148ra116.
202. Halachev MR, Chan JZ-M, Constantinidou CI, Cumley N, Bradley C, Smith-Banks M, et al. Genomic epidemiology of a protracted hospital outbreak caused by multidrug-resistant *Acinetobacter baumannii* in Birmingham, England. *Genome Med.* 2014;6(11):70.

203. Inns T, Lane C, Peters T, Dallman T, Chatt C, McFarland N, et al. A multi-country *Salmonella* Enteritidis phage type 14b outbreak associated with eggs from a German producer: “near real-time” application of whole genome sequencing and food chain investigations, United Kingdom, May to September 2014. *Euro Surveill.* 2015 Apr 23;20(16).
204. Mate SE, Kugelman JR, Nyenswah TG, Ladner JT, Wiley MR, Cordier-Lassalle T, et al. Molecular Evidence of Sexual Transmission of Ebola Virus. *N Engl J Med.* 2015 Dec 17;373(25):2448–54.
205. Reuter S, Török ME, Holden MTG, Reynolds R, Raven KE, Blane B, et al. Building a genomic framework for prospective MRSA surveillance in the United Kingdom and the Republic of Ireland. *Genome Res.* 2016 Feb;26(2):263–70.
206. Santella RM. Approaches to DNA/RNA Extraction and whole genome amplification. *Cancer Epidemiol Biomarkers Prev.* 2006 Sep;15(9):1585–7.
207. Karamitros T, Magiorkinis G. A novel method for the multiplexed target enrichment of MinION next generation sequencing libraries using PCR-generated baits. *Nucleic Acids Res.* 2015 Dec 15;43(22):e152.
208. Thoendel M, Jeraldo PR, Greenwood-Quaintance KE, Yao JZ, Chia N, Hanssen AD, et al. Comparison of microbial DNA enrichment tools for metagenomic whole genome sequencing. *J Microbiol Methods.* 2016;127:141–5.
209. Guo Y, Li N, Lysén C, Frace M, Tang K, Sammons S, et al. Isolation and enrichment of *Cryptosporidium* DNA and verification of DNA purity for whole-genome sequencing. *J Clin Microbiol.* 2015 Feb;53(2):641–7.
210. Sedlackova T, Repiska G, Celec P, Szemes T, Minarik G. Fragmentation of DNA affects the accuracy of the DNA quantitation by the commonly used methods. *Biol Proced Online.* 2013 Feb 13;15(1):5.
211. Nakayama Y, Yamaguchi H, Einaga N, Esumi M. Pitfalls of DNA Quantification Using DNA-Binding Fluorescent Dyes and Suggested Solutions. *PLoS ONE.* 2016;11(3):e0150528.
212. Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, et al. Library construction for next-generation sequencing: overviews and challenges. *BioTechniques.* 2014;56(2):61–4, 66, 68, passim.
213. van Dijk EL, Jaszczyszyn Y, Thermes C. Library preparation methods for next-generation sequencing: tone down the bias. *Exp Cell Res.* 2014 Mar 10;322(1):12–20.
214. Pankhurst LJ, del Ojo Elias C, Votintseva AA, Walker TM, Cole K, Davies J, et al. Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: a prospective study. *Lancet Respir Med.* 2016 Jan;4(1):49–58.

215. Allard MW, Strain E, Melka D, Bunning K, Musser SM, Brown EW, et al. Practical Value of Food Pathogen Traceability through Building a Whole-Genome Sequencing Network and Database. *J Clin Microbiol.* 2016;54(8):1975–83.
216. Ashton PM, Nair S, Peters TM, Bale JA, Powell DG, Painset A, et al. Identification of Salmonella for public health surveillance using whole genome sequencing. *PeerJ.* 2016;4:e1752.
217. Turton JF, Doumith M, Hopkins KL, Perry C, Meunier D, Woodford N. Clonal expansion of Escherichia coli ST38 carrying a chromosomally integrated OXA-48 carbapenemase gene. *J Med Microbiol.* 2016 Jun;65(6):538–46.
218. Quick J, Ashton P, Calus S, Chatt C, Gossain S, Hawker J, et al. Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. *Genome Biol.* 2015 May 30;16:114.
219. Quick J, Quinlan AR, Loman NJ. A reference bacterial genome dataset generated on the MinION™ portable single-molecule nanopore sequencer. *Gigascience.* 2014;3:22.
220. Glenn TC. Field guide to next-generation DNA sequencers: Field Guide to Next-Gen Sequencers. *Molecular Ecology Resources.* 2011 Sep;11(5):759–69.
221. Illumina. Sequencing Platforms [Internet]. [cited 2018 Jan 14]. Available from: <https://www.illumina.com/systems/sequencing-platforms.html>
222. ThermoFisher. IonTorrent PGM Specification Sheet [Internet]. [cited 2018 Jan 14]. Available from: <https://tools.thermofisher.com/content/sfs/brochures/PGM-Specification-Sheet.pdf>
223. ThermoFisher. Ion S5 Specification Sheet [Internet]. [cited 2016 Jul 4]. Available from: <https://www.thermofisher.com/ca/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-workflow/ion-torrent-next-generation-sequencing-run-sequence/ion-s5-ngs-targeted-sequencing/ion-s5-specifications.html>
224. ThermoFisher. Ion Proton Specification Sheet [Internet]. [cited 2018 Jan 14]. Available from: <https://www.thermofisher.com/order/catalog/product/4476610?ICID%20=%20search-product>
225. Oxford Nanopore. MinION Specification Sheet [Internet]. Available from: <https://www.nanoporetech.com/products/specifications>
226. Ip CLC, Loose M, Tyson JR, de Cesare M, Brown BL, Jain M, et al. MinION Analysis and Reference Consortium: Phase 1 data release and analysis. *F1000Res.* 2015;4:1075.
227. PacBio. PacBio Sequencing Systems [Internet]. [cited 2018 Jan 14]. Available from: <http://www.pacb.com/products-and-services/pacbio-systems/>

228. Pightling AW, Petronella N, Pagotto F. Choice of reference-guided sequence assembler and SNP caller for analysis of *Listeria monocytogenes* short-read sequence data greatly influences rates of error. *BMC Res Notes*. 2015 Dec 8;8:748.
229. Illumina. Sequencing coverage calculator [Internet]. [cited 2018 Jan 14]. Available from: https://support.illumina.com/downloads/sequencing_coverage_calculator.html
230. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res*. 2010 Apr;38(6):1767–71.
231. Edwards DJ, Holt KE. Beginner’s guide to comparative bacterial genome analysis using next-generation sequence data. *Microb Inform Exp*. 2013 Apr 10;3(1):2.
232. Holt KE. Tools for bacterial comparative genomics [Internet]. [cited 2018 Jan 14]. Available from: <https://holtlab.net/2015/02/25/tools-for-bacterial-comparative-genomics/>
233. Liao Y-C, Lin S-H, Lin H-H. Completing bacterial genome assemblies: strategy and performance comparisons. *Sci Rep*. 2015 Mar 4;5:8747.
234. Nicholson AC, Whitney AM, Emery BD, Bell ME, Gartin JT, Humrighouse BW, et al. Complete Genome Sequences of Four Strains from the 2015-2016 *Elizabethkingia anophelis* Outbreak. *Genome Announc*. 2016 Jun 16;4(3).
235. Lee RS, Behr MA. Does Choice Matter? Reference-Based Alignment for Molecular Epidemiology of Tuberculosis. *J Clin Microbiol*. 2016;54(7):1891–5.
236. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol*. 2016 Jun 20;17(1):132.
237. Stucki D, Gagneux S. Single nucleotide polymorphisms in *Mycobacterium tuberculosis* and the need for a curated database. *Tuberculosis (Edinb)*. 2013 Jan;93(1):30–9.
238. Mielczarek M, Szyda J. Review of alignment and SNP calling algorithms for next-generation sequencing data. *J Appl Genet*. 2016 Feb;57(1):71–9.
239. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2010 Mar 1;26(5):589–95.
240. Babraham Bioinformatics. FastQC A Quality Control tool for High Throughput Sequence Data [Internet]. [cited 2018 Jan 14]. Available from: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
241. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014 Aug 1;30(15):2114–20.

242. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE*. 2014;9(11):e112963.
243. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010 Sep;20(9):1297–303.
244. Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, et al. A survey of tools for variant analysis of next-generation genome sequencing data. *Brief Bioinformatics*. 2014 Mar;15(2):256–78.
245. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 2011 Nov 1;27(21):2987–93.
246. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. *Bioinformatics*. 2011 Aug 1;27(15):2156–8.
247. Greenfield P, Duesing K, Papanicolaou A, Bauer DC. Blue: correcting sequencing errors using consensus and context. *Bioinformatics*. 2014 Oct;30(19):2723–32.
248. Marçais G, Yorke JA, Zimin A. QuorUM: An Error Corrector for Illumina Reads. *PLoS ONE*. 2015;10(6):e0130821.
249. Olson ND, Lund SP, Colman RE, Foster JT, Sahl JW, Schupp JM, et al. Best practices for evaluating single nucleotide variant calling methods for microbial genomics. *Front Genet*. 2015;6:235.
250. Quinlan AR. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics*. 2014 Sep 8;47:11.12.1-34.
251. Seemann T. snippy: Rapid bacterial SNP calling and core genome alignments [Internet]. 2018. Available from: <https://github.com/tseemann/snippy>
252. Public Health England. PHEnix: Public Health England SNP calling pipeline [Internet]. 2018. Available from: <https://github.com/phe-bioinformatics/PHEnix>
253. Davis S, Pettengill JB, Luo Y, Payne J, Shpuntoff A, Rand H, et al. CFSAN SNP Pipeline: an automated method for constructing SNP matrices from next-generation sequence data. *PeerJ Comput Sci*. 2015 Aug 26;1:e20.
254. Bekal S, Berry C, Reimer AR, Van Domselaar G, Beaudry G, Fournier E, et al. Usefulness of High-Quality Core Genome Single-Nucleotide Variant Analysis for Subtyping the Highly Clonal and the Most Prevalent *Salmonella enterica* Serovar Heidelberg Clone in the Context of Outbreak Investigations. *J Clin Microbiol*. 2016 Feb;54(2):289–95.

255. Wyres KL, Conway TC, Garg S, Queiroz C, Reumann M, Holt K, et al. WGS Analysis and Interpretation in Clinical and Public Health Microbiology Laboratories: What Are the Requirements and How Do Existing Tools Compare? *Pathogens*. 2014 Jun 11;3(2):437–58.
256. Budowle B, Connell ND, Bielecka-Oder A, Colwell RR, Corbett CR, Fletcher J, et al. Validation of high throughput sequencing and microbial forensics applications. *Investig Genet*. 2014;5:9.
257. Foongladda S, Klayut W, Chinli R, Pholwat S, Houpt ER. Use of mycobacteriophage quantitative PCR on MGIT broths for a rapid tuberculosis antibiogram. *J Clin Microbiol*. 2014 May;52(5):1523–8.
258. Simons SO, van Soolingen D. Drug susceptibility testing for optimizing tuberculosis treatment. *Curr Pharm Des*. 2011;17(27):2863–74.
259. Yakrus MA, Driscoll J, Lentz AJ, Sikes D, Hartline D, Metchock B, et al. Concordance between molecular and phenotypic testing of *Mycobacterium tuberculosis* complex isolates for resistance to rifampin and isoniazid in the United States. *J Clin Microbiol*. 2014 Jun;52(6):1932–7.
260. Banu S, Rahman SMM, Khan MSR, Ferdous SS, Ahmed S, Gratz J, et al. Discordance across several methods for drug susceptibility testing of drug-resistant *Mycobacterium tuberculosis* isolates in a single laboratory. *J Clin Microbiol*. 2014 Jan;52(1):156–63.
261. Centers for Disease Control and Prevention. Report of Expert Consultations on Rapid Molecular Testing to Detect Drug-Resistant Tuberculosis in the United States [Internet]. 2009 [cited 2018 Jan 14]. Available from: <https://www.cdc.gov/tb/topic/laboratory/rapidmoleculartesting/moldstreport.pdf>
262. Nebenzahl-Guimaraes H, Jacobson KR, Farhat MR, Murray MB. Systematic review of allelic exchange experiments aimed at identifying mutations that confer drug resistance in *Mycobacterium tuberculosis*. *J Antimicrob Chemother*. 2014 Feb;69(2):331–42.
263. Morlock GP, Crawford JT, Butler WR, Brim SE, Sikes D, Mazurek GH, et al. Phenotypic characterization of *pncA* mutants of *Mycobacterium tuberculosis*. *Antimicrob Agents Chemother*. 2000 Sep;44(9):2291–5.
264. Alexander DC, Ma JH, Guthrie JL, Blair J, Chedore P, Jamieson FB. Gene sequencing for routine verification of pyrazinamide resistance in *Mycobacterium tuberculosis*: a role for *pncA* but not *rpsA*. *Journal of clinical microbiology*. 2012;50(11):3726–3728.
265. Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, et al. Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. *Nat Genet*. 2013 Oct;45(10):1183–9.

266. Coll F, McNerney R, Preston MD, Guerra-Assunção JA, Warry A, Hill-Cawthorne G, et al. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. *Genome Med.* 2015;7(1):51.
267. Walker TM, Kohl TA, Omar SV, Hedge J, Del Ojo Elias C, Bradley P, et al. Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: a retrospective cohort study. *Lancet Infect Dis.* 2015 Oct;15(10):1193–202.
268. Farhat MR, Sultana R, Iartchouk O, Bozeman S, Galagan J, Sisk P, et al. Genetic Determinants of Drug Resistance in *Mycobacterium tuberculosis* and Their Diagnostic Value. *Am J Respir Crit Care Med.* 2016 Sep 1;194(5):621–30.
269. Schito M, Dolinger DL. A Collaborative Approach for “ReSeq-ing” *Mycobacterium tuberculosis* Drug Resistance: Convergence for Drug and Diagnostic Developers. *EBioMedicine.* 2015 Oct;2(10):1262–5.
270. Hillemann D, Rüscher-Gerdes S, Richter E. Evaluation of the GenoType MTBDRplus assay for rifampin and isoniazid susceptibility testing of *Mycobacterium tuberculosis* strains and clinical specimens. *J Clin Microbiol.* 2007 Aug;45(8):2635–40.
271. Rossau R, Traore H, De Beenhouwer H, Mijs W, Jannes G, De Rijk P, et al. Evaluation of the INNO-LiPA Rif. TB assay, a reverse hybridization assay for the simultaneous detection of *Mycobacterium tuberculosis* complex and its resistance to rifampin. *Antimicrob Agents Chemother.* 1997 Oct;41(10):2093–8.
272. Steiner A, Stucki D, Coscolla M, Borrell S, Gagneux S. KvarQ: targeted and direct variant calling from fastq reads of bacterial genomes. *BMC Genomics.* 2014 Oct 9;15:881.
273. Feuerriegel S, Schleusener V, Beckert P, Kohl TA, Miotto P, Cirillo DM, et al. PhyResSE: a Web Tool Delineating *Mycobacterium tuberculosis* Antibiotic Resistance and Lineage from Whole-Genome Sequencing Data. *J Clin Microbiol.* 2015 Jun;53(6):1908–14.
274. Bradley P, Gordon NC, Walker TM, Dunn L, Heys S, Huang B, et al. Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*. *Nat Commun.* 2015 Dec 21;6:10063.
275. Davies J, Davies D. Origins and evolution of antibiotic resistance. *Microbiol Mol Biol Rev.* 2010 Sep;74(3):417–33.
276. Levy SB, Marshall B. Antibacterial resistance worldwide: causes, challenges and responses. *Nat Med.* 2004 Dec;10(12 Suppl):S122-129.
277. Inouye M, Dashnow H, Raven L-A, Schultz MB, Pope BJ, Tomita T, et al. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med.* 2014;6(11):90.

278. Rowe W, Baker KS, Verner-Jeffreys D, Baker-Austin C, Ryan JJ, Maskell D, et al. Search Engine for Antimicrobial Resistance: A Cloud Compatible Pipeline and Web Interface for Rapidly Detecting Antimicrobial Resistance Genes Directly from Sequence Data. *PLoS ONE*. 2015;10(7):e0133492.
279. Hudson CM, Bent ZW, Meagher RJ, Williams KP. Resistance determinants and mobile genetic elements of an NDM-1-encoding *Klebsiella pneumoniae* strain. *PLoS ONE*. 2014;9(6):e99209.
280. Adam M, Murali B, Glenn NO, Potter SS. Epigenetic inheritance based evolution of antibiotic resistance in bacteria. *BMC Evol Biol*. 2008 Feb 18;8:52.
281. Salinas-Delgado Y, Galaviz-Hernández C, Toral RG, Ávila Rejón CA, Reyes-Lopez MA, Martínez AR, et al. The D543N polymorphism of the SLC11A1/NRAMP1 gene is associated with treatment failure in male patients with pulmonary tuberculosis. *Drug Metab Pers Ther*. 2015 Sep;30(3):211–4.
282. Alexander DC, Fitzgerald SF, DePaulo R, Kitzul R, Daku D, Levett PN, et al. Laboratory-Acquired Infection with *Salmonella enterica* Serovar Typhimurium Exposed by Whole-Genome Sequencing. *J Clin Microbiol*. 2016 Jan;54(1):190–3.
283. Moore G, Cookson B, Gordon NC, Jackson R, Kearns A, Singleton J, et al. Whole-genome sequencing in hierarchy with pulsed-field gel electrophoresis: the utility of this approach to establish possible sources of MRSA cross-transmission. *J Hosp Infect*. 2015 May;90(1):38–45.
284. Mac Aogáin M, Moloney G, Kilkenny S, Kelleher M, Kelleghan M, Boyle B, et al. Whole-genome sequencing improves discrimination of relapse from reinfection and identifies transmission events among patients with recurrent *Clostridium difficile* infections. *J Hosp Infect*. 2015 Jun;90(2):108–16.
285. Abecasis AB, Geretti AM, Albert J, Power L, Weait M, Vandamme A-M. Science in court: the myth of HIV fingerprinting. *Lancet Infect Dis*. 2011 Feb;11(2):78–9.
286. Jombart T, Aanensen DM, Baguelin M, Birrell P, Cauchemez S, Camacho A, et al. OutbreakTools: a new platform for disease outbreak analysis using the R software. *Epidemics*. 2014 Jun;7:28–34.
287. Worby CJ, Read TD. “SEEDY” (Simulation of Evolutionary and Epidemiological Dynamics): An R Package to Follow Accumulation of Within-Host Mutation in Pathogens. *PLoS ONE*. 2015;10(6):e0129745.
288. Worby CJ, Lipsitch M, Hanage WP. Within-host bacterial diversity hinders accurate reconstruction of transmission networks from genomic distance data. *PLoS Comput Biol*. 2014 Mar;10(3):e1003549.

289. Quick J, Cumley N, Wearn CM, Niebel M, Constantinidou C, Thomas CM, et al. Seeking the source of *Pseudomonas aeruginosa* infections in a recently opened hospital: an observational study using whole-genome sequencing. *BMJ Open*. 2014 Nov 4;4(11):e006278.
290. Hoffmann M, Luo Y, Monday SR, Gonzalez-Escalona N, Ottesen AR, Muruvanda T, et al. Tracing Origins of the *Salmonella* Bareilly Strain Causing a Food-borne Outbreak in the United States. *J Infect Dis*. 2016 Feb 15;213(4):502–8.
291. Jackson BR, Tarr C, Strain E, Jackson KA, Conrad A, Carleton H, et al. Implementation of Nationwide Real-time Whole-genome Sequencing to Enhance Listeriosis Outbreak Detection and Investigation. *Clin Infect Dis*. 2016 Aug 1;63(3):380–6.
292. Bryant JM, Grogono DM, Greaves D, Foweraker J, Roddick I, Inns T, et al. Whole-genome sequencing to identify transmission of *Mycobacterium abscessus* between patients with cystic fibrosis: a retrospective cohort study. *Lancet*. 2013 May 4;381(9877):1551–60.
293. Holt KE, Baker S, Weill F-X, Holmes EC, Kitchen A, Yu J, et al. *Shigella sonnei* genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. *Nat Genet*. 2012 Sep;44(9):1056–9.
294. Golubchik T, Brueggemann AB, Street T, Gertz RE, Spencer CCA, Ho T, et al. Pneumococcal genome sequencing tracks a vaccine escape variant formed through a multi-fragment recombination event. *Nat Genet*. 2012 Jan 29;44(3):352–5.
295. Grad YH, Kirkcaldy RD, Trees D, Dordel J, Harris SR, Goldstein E, et al. Genomic epidemiology of *Neisseria gonorrhoeae* with reduced susceptibility to cefixime in the USA: a retrospective observational study. *Lancet Infect Dis*. 2014 Mar;14(3):220–6.
296. Loman NJ, Pallen MJ. Twenty years of bacterial genome sequencing. *Nat Rev Microbiol*. 2015;13(12):787–94.
297. Grad YH, Lipsitch M. Epidemiologic data and pathogen genome sequences: a powerful synergy for public health. *Genome Biol*. 2014 Nov 18;15(11):538.
298. World Health Organization. Global strategy and targets for tuberculosis prevention, care and control after 2015 [Internet]. Geneva, Switzerland; 2014. Report No.: A67/11. Available from: http://apps.who.int/gb/ebwha/pdf_files/WHA67/A67_11-en.pdf
299. Wlodarska M, Johnston JC, Gardy JL, Tang P. A microbiological revolution meets an ancient disease: improving the management of tuberculosis with genomics. *Clin Microbiol Rev*. 2015 Apr;28(2):523–39.
300. Galagan JE. Genomic insights into tuberculosis. *Nat Rev Genet*. 2014 May;15(5):307–20.

301. Hodkinson BP, Grice EA. Next-Generation Sequencing: A Review of Technologies and Tools for Wound Microbiome Research. *Adv Wound Care (New Rochelle)*. 2015 Jan 1;4(1):50–8.
302. Gardy J, Loman N, Underwood A, Schaik W van, Connor T, Seemann T, et al. ABPHM15 EtherPad Archive [Internet]. 2015 [cited 2015 May 16]. Available from: http://figshare.com/articles/ABPHM15_EtherPad_Archive/1408783
303. Louw GE, Warren RM, Gey van Pittius NC, McEvoy CRE, Van Helden PD, Victor TC. A balancing act: efflux/influx in mycobacterial drug resistance. *Antimicrob Agents Chemother*. 2009 Aug;53(8):3181–9.
304. Köser CU, Bryant JM, Becq J, Török ME, Ellington MJ, Marti-Renom MA, et al. Whole-Genome Sequencing for Rapid Susceptibility Testing of *M. tuberculosis*. *New England Journal of Medicine*. 2013 Jul 18;369(3):290–2.
305. Adam HJ, Guthrie JL, Bolotin S, Alexander DC, Stuart R, Pyskir D, et al. Genotypic characterization of tuberculosis transmission within Toronto’s under-housed population, 1997–2008. *The International Journal of Tuberculosis and Lung Disease*. 2010;14(10):1350–1353.
306. de Beer JL, Ingen J van, Vries G de, Erkens C, Sebek M, Mulder A, et al. Comparative Study of IS6110 Restriction Fragment Length Polymorphism and Variable-Number Tandem-Repeat Typing of *Mycobacterium tuberculosis* Isolates in the Netherlands, Based on a 5-Year Nationwide Survey. *J Clin Microbiol*. 2013 Apr 1;51(4):1193–8.
307. Guthrie JL, Alexander DC, Marchand-Austin A, Lam K, Whelan M, Lee B, et al. Technology and tuberculosis control: the OUT-TB Web experience. *J Am Med Inform Assoc*. 2017 Apr 1;24(e1):e136–42.
308. Centers for Disease Control and Prevention. New CDC program for rapid genotyping of *Mycobacterium tuberculosis* isolates. *MMWR*. 2005 Jan;54(2):47.
309. Bauer J, Kok-Jensen A, Faurschou P, Thuesen J, Taudorf E, Andersen AB. A prospective evaluation of the clinical value of nation-wide DNA fingerprinting of tuberculosis isolates in Denmark. *Int J Tuberc Lung Dis*. 2000 Apr;4(4):295–9.
310. Bidovec-Stojkovic U, Zolnir-Dovc M, Supply P. One year nationwide evaluation of 24-locus MIRU-VNTR genotyping on Slovenian *Mycobacterium tuberculosis* isolates. *Respiratory Medicine*. 2011 Oct 1;105:S67–73.
311. van Soolingen D, Borgdorff MW, de Haas PE, Sebek MM, Veen J, Dessens M, et al. Molecular epidemiology of tuberculosis in the Netherlands: a nationwide study from 1993 through 1997. *J Infect Dis*. 1999 Sep;180(3):726–36.
312. National TB strain typing service: what we do [Internet]. [cited 2017 Jul 7]. Available from: <https://www.gov.uk/guidance/national-tb-strain-typing-service-what-we-do>

313. Clark CM, Driver CR, Munsiff SS, Driscoll JR, Kreiswirth BN, Zhao B, et al. Universal Genotyping in Tuberculosis Control Program, New York City, 2001–2003. *Emerging Infectious Diseases*. 2006 May;12(5):719–24.
314. Lambregts-van Weezenbeek CSB, Sebek MMGG, van Gerven PJHJ, de Vries G, Verver S, Kalisvaart NA, et al. Tuberculosis contact investigation and DNA fingerprint surveillance in The Netherlands: 6 years' experience with nation-wide cluster feedback and cluster monitoring. *The International Journal of Tuberculosis and Lung Disease*. 2003 Dec 1;7(12):S463–70.
315. Reves R. Universal Genotyping as a Tool for Establishing Successful Partnerships for Tuberculosis Elimination. *American Journal of Respiratory and Critical Care Medicine*. 2006 Sep;174(5):491–2.
316. Matheson FI, Dunn JR, Smith KLW, Moineddin R, Glazier RH. Development of the Canadian Marginalization Index: a new tool for the study of inequality. *Can J Public Health*. 2012 Apr 30;103(8 Suppl 2):S12-16.
317. Oeltmann JE, Kammerer JS, Pevzner ES, Moonan PK. Tuberculosis and Substance Abuse in the United States, 1997-2006. *Arch Intern Med*. 2009 Jan 26;169(2):189–97.
318. Comas I, Homolka S, Niemann S, Gagneux S. Genotyping of Genetically Monomorphic Bacteria: DNA Sequencing in *Mycobacterium tuberculosis* Highlights the Limitations of Current Methodologies. *PLoS One*. 2009 Nov 12;4(11):e7815.
319. Ghosh S, Moonan PK, Cowan L, Grant J, Kammerer S, Navin TR. Tuberculosis Genotyping Information Management System: Enhancing Tuberculosis Surveillance in the United States. *Infection, Genetics and Evolution*. 2012 Jun 1;12(4):782–8.
320. Northrup JM, Miller AC, Nardell E, Sharnprapai S, Etkind S, Driscoll J, et al. Estimated Costs of False Laboratory Diagnoses of Tuberculosis in Three Patients. *Emerg Infect Dis*. 2002 Nov;8(11):1264–70.
321. Cook VJ, Stark G, Roscoe DL, Kwong A, Elwood RK. Investigation of suspected laboratory cross-contamination: interpretation of single smear-negative, positive cultures for *Mycobacterium tuberculosis*. *Clinical Microbiology and Infection*. 2006 Oct 1;12(10):1042–5.
322. Hawkey PM, Smith EG, Evans JT, Monk P, Bryan G, Mohamed HH, et al. Mycobacterial Interspersed Repetitive Unit Typing of *Mycobacterium tuberculosis* Compared to IS6110-Based Restriction Fragment Length Polymorphism Analysis for Investigation of Apparently Clustered Cases of Tuberculosis. *J Clin Microbiol*. 2003 Aug 1;41(8):3514–20.

323. Statistics Canada, Government of Canada. Vancouver, CMA. NHS Focus on Geography Series. [Internet]. 2013 [cited 2017 Jun 1]. Available from: <http://www12.statcan.gc.ca/nhs-enm/2011/as-sa/fogs-spg/Pages/FOG.cfm?lang=E&level=3&GeoCode=933>
324. Hernández-Garduño E, Kunimoto D, Wang L, Rodrigues M, Elwood RK, Black W, et al. Predictors of clustering of tuberculosis in Greater Vancouver: a molecular epidemiologic study. *CMAJ*. 2002 Aug 20;167(4):349–52.
325. Kulaga S, Behr M, Musana K, Brinkman J, Menzies D, Brassard P, et al. Molecular epidemiology of tuberculosis in Montreal. *CMAJ*. 2002 Aug 20;167(4):353–4.
326. Alexander DC, Guthrie JL, Pyskir D, Maki A, Kurepina N, Kreiswirth BN, et al. Mycobacterium tuberculosis in Ontario, Canada: insights from IS6110 restriction fragment length polymorphism and mycobacterial interspersed repetitive-unit-variable-number tandem-repeat genotyping. *Journal of clinical microbiology*. 2009;47(8):2651–2654.
327. Blackwood KS, Al-Azem A, Elliott LJ, Hershfield ES, Kabani AM. Conventional and molecular epidemiology of tuberculosis in Manitoba. *BMC Infect Dis*. 2003 Aug 13;3:18.
328. FitzGerald J. M, Fanning A, Hoepfner V, Hershfield E, Kunimoto D, Canadian Molecular Epidemiology of TB Study Group. The molecular epidemiology of tuberculosis in Western Canada. *The International Journal of Tuberculosis and Lung Disease*. 2003 Feb 1;7(2):132–8.
329. Shabbeer A, Cowan LS, Ozcaglar C, Rastogi N, Vandenberg SL, Yener B, et al. TB-Lineage: an online tool for classification and analysis of strains of Mycobacterium tuberculosis complex. *Infect Genet Evol*. 2012 Jun;12(4):789–97.
330. Francisco AP, Vaz C, Monteiro PT, Melo-Cristino J, Ramirez M, Carriço JA. PHYLOViZ: phylogenetic inference and data visualization for sequence based typing methods. *BMC Bioinformatics*. 2012;13:87.
331. Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*. 1974 Dec;19(6):716–23.
332. Little RJA. A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association*. 1988 Dec 1;83(404):1198–202.
333. World Health Organization. Global tuberculosis report 2016 [Internet]. WHO. [cited 2018 Feb 1]. Available from: http://www.who.int/tb/publications/global_report/en/
334. Griffith A. *Multiculturalism in Canada: Evidence and Anecdote*. Ottawa, Ontario: Anar Press; 2015. 370 p.

335. de Jong BC, Hill PC, Aiken A, Awine T, Antonio M, Adetifa IM, et al. Progression to active tuberculosis, but not transmission, varies by *M. tuberculosis* lineage in The Gambia. *J Infect Dis*. 2008 Oct 1;198(7):1037–43.
336. Stucki D, Ballif M, Egger M, Furrer H, Altpeter E, Battegay M, et al. Standard Genotyping Overestimates Transmission of *Mycobacterium tuberculosis* among Immigrants in a Low-Incidence Country. *J Clin Microbiol*. 2016 Jul 1;54(7):1862–70.
337. Geng E, Kreiswirth B, Driver C, Li J, Burzynski J, DellaLatta P, et al. Changes in the Transmission of Tuberculosis in New York City from 1990 to 1999. *New England Journal of Medicine*. 2002 May 9;346(19):1453–8.
338. Sandgren A, Sañé Schepisi M, Sotgiu G, Huitric E, Migliori GB, Manissero D, et al. Tuberculosis transmission between foreign- and native-born populations in the EU/EEA: a systematic review. *Eur Respir J*. 2014 Apr;43(4):1159–71.
339. Lofy KH, McElroy PD, Lake L, Cowan LS, Diem LA, Goldberg SV, et al. Outbreak of tuberculosis in a homeless population involving multiple sites of transmission. *The International Journal of Tuberculosis and Lung Disease*. 2006 Jun 1;10(6):683–9.
340. Tan de Bibiana J, Rossi C, Rivest P, Zwerling A, Thibert L, McIntosh F, et al. Tuberculosis and homelessness in Montreal: a retrospective cohort study. *BMC Public Health*. 2011;11:833.
341. Merker M, Blin C, Mona S, Duforet-Frebourg N, Lecher S, Willery E, et al. Evolutionary history and global spread of the *Mycobacterium tuberculosis* Beijing lineage. *Nat Genet*. 2015 Mar;47(3):242–9.
342. Anderson LF, Tamne S, Brown T, Watson JP, Mullarkey C, Zenner D, et al. Transmission of multidrug-resistant tuberculosis in the UK: a cross-sectional molecular and epidemiological study of clustering and contact tracing. *The Lancet Infectious Diseases*. 2014 May;14(5):406–15.
343. Roetzer A, Schuback S, Diel R, Gasau F, Ubben T, Nauta A di, et al. Evaluation of *Mycobacterium tuberculosis* Typing Methods in a 4-Year Study in Schleswig-Holstein, Northern Germany. *J Clin Microbiol*. 2011 Dec 1;49(12):4173–8.
344. Colijn C, Cohen T, Murray M. Emergent heterogeneity in declining tuberculosis epidemics. *Journal of Theoretical Biology*. 2007 Aug;247(4):765–74.
345. Long R, Heffernan C, Gao Z, Egedahl ML, Talbot J. Do “Virtual” and “Outpatient” Public Health Tuberculosis Clinics Perform Equally Well? A Program-Wide Evaluation in Alberta, Canada. *PLoS ONE*. 2015;10(12):e0144784.

346. Sloom R, Borgdorff MW, Beer JL de, Ingen J van, Supply P, Soolingen D van. Clustering of Tuberculosis Cases Based on Variable-Number Tandem-Repeat Typing in Relation to the Population Structure of Mycobacterium tuberculosis in the Netherlands. *J Clin Microbiol.* 2013 Jul 1;51(7):2427–31.
347. Murdie RA, University of Toronto, Cities Centre. Diversity and concentration in Canadian immigration: trends in Toronto, Montréal and Vancouver, 1971-2006 [Internet]. Toronto, Ont.: Centre for Urban & Community Studies; 2011 [cited 2018 Jul 15]. Available from: <http://www.deslibris.ca/ID/226201>
348. Statistics Canada. Annual Demographic Estimates: Canada, Provinces and Territories. Catalogue no. 91-215-X [Internet]. 2017 Sep. Available from: <https://www150.statcan.gc.ca/n1/pub/91-215-x/91-215-x2017000-eng.pdf>
349. Eberle MP. Homelessness, Causes & Effects, Vol. 2: A Profile, Policy Review and Analysis of Homelessness in British Columbia. Victoria, B.C.: Ministry of Social Development and Economic Security; 2001.
350. Allidina A, Humphrey M, John K. Linking Services to Outcomes: A Report on the Relationship between Child Welfare Services and Youth Homelessness in Canada for the YWCA Canada [Internet]. 2015 [cited 2018 Jun 21]. Available from: http://ywcacanada.ca/data/research_docs/00000348.pdf
351. Statistics Canada. Canadian Demographics at a Glance, Second edition [Internet]. 2016 [cited 2018 Jun 16]. Available from: <http://www.statcan.gc.ca/pub/91-003-x/91-003-x2014001-eng.htm>
352. Government of Canada SC. 2014 Health Region Peer Groups – User Guide [Internet]. 2015 [cited 2018 Jun 19]. Available from: <http://www.statcan.gc.ca/pub/82-402-x/2015001/regions/hrpg2014-eng.htm>
353. Menéndez MC, Buxton RS, Evans JT, Gascoyne-Binzi D, Barlow REL, Hinds J, et al. Genome analysis shows a common evolutionary origin for the dominant strains of Mycobacterium tuberculosis in a UK South Asian community. *Tuberculosis.* 2007 Sep 1;87(5):426–36.
354. Wyllie D, Davidson J, Walker T, Rathod P, Peto T, Robinson E, et al. A quantitative evaluation of MIRU-VNTR typing against whole-genome sequencing for identifying Mycobacterium tuberculosis transmission: A prospective observational cohort study. *bioRxiv.* 2018 Jan 24;252734.
355. Jonsson J, Hoffner S, Berggren I, Bruchfeld J, Ghebremichael S, Pennhag A, et al. Comparison between RFLP and MIRU-VNTR Genotyping of Mycobacterium tuberculosis Strains Isolated in Stockholm 2009 to 2011. *PLOS ONE.* 2014 Apr 14;9(4):e95159.

356. Mitruka K, Blake H, Ricks P, Miramontes R, Bamrah S, Chee C, et al. A Tuberculosis Outbreak Fueled by Cross-Border Travel and Illicit Substances: Nevada and Arizona. *Public Health Rep.* 2014;129(1):78–85.
357. Fiebig L, Kohl TA, Popovici O, Mühlenfeld M, Indra A, Homorodean D, et al. A joint cross-border investigation of a cluster of multidrug-resistant tuberculosis in Austria, Romania and Germany in 2014 using classic, genotyping and whole genome sequencing methods: lessons learnt. *Eurosurveillance* [Internet]. 2017 Jan 12;22(2). Available from: <http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=22686>
358. Lathan M, Mukasa LN, Hooper N, Golub J, Baruch N, Mulcahy D, et al. Cross-Jurisdictional Transmission of Mycobacterium tuberculosis in Maryland and Washington, D.C., 1996-2000, Linked to the Homeless. *Emerging Infectious Diseases.* 2002 Nov;8(11):1249–51.
359. Garzelli C, Rindi L. Molecular epidemiological approaches to study the epidemiology of tuberculosis in low-incidence settings receiving immigrants. *Infection, Genetics and Evolution.* 2012 Jun;12(4):610–8.
360. Marais BJ. Childhood Tuberculosis: Epidemiology and Natural History of Disease. *Indian J Pediatr.* 2011 Mar 1;78(3):321–7.
361. Brent AJ, Anderson ST, Kampmann B. Childhood tuberculosis: out of sight, out of mind? *Trans R Soc Trop Med Hyg.* 2008 Mar;102(3–4):217–8.
362. Sun SJ, Bennett DE, Flood J, Loeffler AM, Kammerer S, Ellis BA. Identifying the Sources of Tuberculosis in Young Children: A Multistate Investigation. *Emerging Infectious Diseases.* 2002 Nov;8(11):1216–23.
363. Sánchez-Albisua I, Baquero-Artigao F, Del Castillo F, Borque C, García-Miguel MJ, Vidal ML. Twenty years of pulmonary tuberculosis in children: what has changed? *Pediatr Infect Dis J.* 2002 Jan;21(1):49–53.
364. Lobato MN, Mohle-Boetani JC, Royce SE. Missed Opportunities for Preventing Tuberculosis Among Children Younger Than Five Years of Age. *Pediatrics.* 2000 Dec 1;106(6):e75–e75.
365. Rayment JH, Guthrie JL, Lam K, Whelan M, Lee B, Jamieson FB, et al. Culture-positive Pediatric Tuberculosis in Toronto, Ontario: Sources of Infection and Relationship of Birthplace and Mycobacterial Lineage to Phenotype. *Pediatr Infect Dis J.* 2016 Jan;35(1):13–8.
366. Schaaf HS, Michaelis IA, Richardson M, Booyesen CN, Gie RP, Warren R, et al. Adult-to-child transmission of tuberculosis: household or community contact? *Int J Tuberc Lung Dis.* 2003 May;7(5):426–31.

367. Wootton SH, Gonzalez BE, Pawlak R, Teeter LD, Smith KC, Musser JM, et al. Epidemiology of Pediatric Tuberculosis Using Traditional and Molecular Techniques: Houston, Texas. *Pediatrics*. 2005 Nov 1;116(5):1141–7.
368. Winston CA, Menzies HJ. Pediatric and Adolescent Tuberculosis in the United States, 2008-2010. *PEDIATRICS*. 2012 Dec 1;130(6):e1425–32.
369. Santiago B, Baquero-Artigao F, Mejías A, Blázquez D, Jiménez MS, Mellado-Peña MJ. Pediatric Drug-resistant Tuberculosis in Madrid: Family Matters. *The Pediatric Infectious Disease Journal*. 2014 Apr;33(4):345–50.
370. Schaaf HS, Van Rie A, Gie RP, Beyers N, Victor TC, Van Helden PD, et al. Transmission of multidrug-resistant tuberculosis: The Pediatric Infectious Disease Journal. 2000 Aug;19(8):695–700.
371. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 2014 May 1;30(9):1312–3.
372. Stucki D, Malla B, Hostettler S, Huna T, Feldmann J, Yeboah-Manu D, et al. Two new rapid SNP-typing methods for classifying Mycobacterium tuberculosis complex into the main phylogenetic lineages. *PLoS ONE*. 2012;7(7):e41253.
373. Votintseva AA, Bradley P, Pankhurst L, Elias C del O, Loose M, Nilgiriwala K, et al. Same-Day Diagnostic and Surveillance Data for Tuberculosis via Whole-Genome Sequencing of Direct Respiratory Samples. *J Clin Microbiol*. 2017 May 1;55(5):1285–98.
374. Saiman L, Gabriel PS, Schulte J, Vargas MP, Kenyon T, Onorato I. Risk Factors for Latent Tuberculosis Infection Among Children in New York City. *Pediatrics*. 2001 May 1;107(5):999–1003.
375. Kik SV, Mensen M, Beltman M, Gijsberts M, van Ameijden EJC, Cobelens FGJ, et al. Risk of travelling to the country of origin for tuberculosis among immigrants living in a low-incidence country. *Int J Tuberc Lung Dis*. 2011 Jan;15(1):38–43.
376. State of Alaska. Tuberculosis in Alaska, 2016 Annual Report [Internet]. 2017 Oct. Available from: http://dhss.alaska.gov/dph/Epi/id/SiteAssets/Pages/TB/TB_Report_2016.pdf
377. Dehghani K, Lan Z, Li P, Michelsen SW, Waites S, Benedetti A, et al. Determinants of tuberculosis trends in six Indigenous populations of the USA, Canada, and Greenland from 1960 to 2014: a population-based study. *The Lancet Public Health*. 2018 Mar 1;3(3):e133–42.
378. Lalor MK, Anderson LF, Hamblion EL, Burkitt A, Davidson JA, Maguire H, et al. Recent household transmission of tuberculosis in England, 2010–2012: retrospective national cohort study combining epidemiological and molecular strain typing data. *BMC Med*. 2017 Jun 13;15(1):105.

379. Bjorn-Mortensen K, Soborg B, Koch A, Ladefoged K, Merker M, Lillebaek T, et al. Tracing Mycobacterium tuberculosis transmission by whole genome sequencing in a high incidence setting: a retrospective population-based study in East Greenland. *Sci Rep*. 2016 Sep 12;6:33180.
380. Government of Canada SC. Statistics Canada: 2011 Census Profile [Internet]. 2012 [cited 2018 Oct 3]. Available from: <http://www12.statcan.gc.ca/census-recensement/2011/dp-pd/prof/details/Page.cfm?Lang=E&Geo1=PR&Code1=60&Geo2=PR&Code2=01&Data=Count&SearchText=Yukon&SearchType=Begins&SearchPR=01&B1=All&GeoLevel=PR&GeoCode=60>
381. Nava-Aguilera E, Andersson N, Harris E, Mitchell S, Hamel C, Shea B, et al. Risk factors associated with recent transmission of tuberculosis: systematic review and meta-analysis. *Int J Tuberc Lung Dis*. 2009 Jan;13(1):17–26.
382. Stein RA. Super-spreaders in infectious diseases. *International Journal of Infectious Diseases*. 2011 Aug 1;15(8):e510–3.
383. MacIntyre CR, Plant AJ, Hulls J, Streeton JA, Graham NMH, Rouch GJ. High Rate of Transmission of Tuberculosis in an Office: Impact of Delayed Diagnosis. *Clinical Infectious Diseases*. 1995;21(5):1170–4.
384. France AM, Grant J, Kammerer JS, Navin TR. A Field-Validated Approach Using Surveillance and Genotyping Data to Estimate Tuberculosis Attributable to Recent Transmission in the United States. *Am J Epidemiol*. 2015 Nov 1;182(9):799–807.
385. Ford CB, Lin PL, Chase MR, Shah RR, Iartchouk O, Galagan J, et al. Use of whole genome sequencing to estimate the mutation rate of Mycobacterium tuberculosis during latent infection. *Nat Genet*. 2011 May;43(5):482–6.
386. Colangeli R, Arcus VL, Cursons RT, Ruthe A, Karalus N, Coley K, et al. Whole genome sequencing of Mycobacterium tuberculosis reveals slow growth and low mutation rates during latent infections in humans. *PLoS ONE*. 2014;9(3):e91024.
387. McEvoy CRE, Cloete R, Müller B, Schürch AC, van Helden PD, Gagneux S, et al. Comparative analysis of Mycobacterium tuberculosis *pe* and *ppe* genes reveals high sequence variation and an apparent absence of selective constraints. *PLoS ONE*. 2012;7(4):e30593.
388. Bainomugisa A, Duarte T, Lavu E, Pandey S, Coulter C, Marais BJ, et al. A complete high-quality MinION nanopore assembly of an extensively drug-resistant Mycobacterium tuberculosis Beijing lineage strain identifies novel variation in repetitive PE/PPE gene regions. *Microb Genom*. 2018 Jun 15;

389. Nguyen D, Proulx J-F, Westley J, Thibert L, Dery S, Behr MA. Tuberculosis in the Inuit Community of Quebec, Canada. *American Journal of Respiratory and Critical Care Medicine*. 2003 Dec;168(11):1353–7.
390. Yukon Bureau of Statistics. Population Report - First Quarter, 2018. Info sheet no. 60 [Internet]. 2018 Jul [cited 2018 Sep 18]. Available from: http://www.eco.gov.yk.ca/stats/pdf/populationQ1_2018.pdf
391. Boyatzis RE. *Transforming Qualitative Information: Thematic Analysis and Code Development*. SAGE; 1998. 204 p.
392. Altman DG. *Practical statistics for medical research*. London: Chapman and Hall; 1991. 611 p.
393. Oelemann MC, Diel R, Vatin V, Haas W, Rüscher-Gerdes S, Loch C, et al. Assessment of an Optimized Mycobacterial Interspersed Repetitive- Unit-Variable-Number Tandem-Repeat Typing System Combined with Spoligotyping for Population-Based Molecular Epidemiology Studies of Tuberculosis. *J Clin Microbiol*. 2007 Mar 1;45(3):691–7.
394. Casali N, Broda A, Harris SR, Parkhill J, Brown T, Drobniowski F. Whole Genome Sequence Analysis of a Large Isoniazid-Resistant Tuberculosis Outbreak in London: A Retrospective Observational Study. *PLOS Medicine*. 2016 Oct 4;13(10):e1002137.
395. Kamper-Jørgensen Z, Andersen AB, Kok-Jensen A, Bygbjerg IC, Andersen PH, Thomsen VO, et al. Clustered Tuberculosis in a Low-Burden Country: Nationwide Genotyping through 15 Years. *J Clin Microbiol*. 2012 Aug;50(8):2660–7.
396. Jajou R, Neeling A de, Rasmussen EM, Norman A, Mulder A, Hunen R van, et al. A Predominant Variable-Number Tandem-Repeat Cluster of Mycobacterium tuberculosis Isolates among Asylum Seekers in the Netherlands and Denmark, Deciphered by Whole-Genome Sequencing. *J Clin Microbiol*. 2018 Feb 1;56(2):e01100-17.
397. Jajou R, Neeling A de, Hunen R van, Vries G de, Schimmel H, Mulder A, et al. Epidemiological links between tuberculosis cases identified twice as efficiently by whole genome sequencing than conventional molecular typing: A population-based study. *PLOS ONE*. 2018 Apr 4;13(4):e0195413.
398. Guthrie JL, Delli Pizzi A, Roth D, Kong C, Jorgensen D, Rodrigues M, et al. Genotyping and Whole Genome Sequencing to Identify Tuberculosis Transmission to Pediatric Patients in British Columbia, Canada, 2005–2014. *The Journal of Infectious Diseases*. 2018 May 11;218(7):1155–1163.
399. Clinical and Laboratory Standards Institute (CLSI): Susceptibility testing of mycobacteria, nocardiae, and other aerobic actinomycetes; approved standard. M24-A2. 2nd ed. Wayne, PA, USA: CLSI; 2011.

400. Menzies D. Molecular methods for tuberculosis trials: time for whole-genome sequencing? *The Lancet Respiratory Medicine*. 2013 Dec 1;1(10):759–61.
401. Maguire H, Brailsford S, Carless J, Yates M, Altass L, Yates S, et al. Large outbreak of isoniazid-monoresistant tuberculosis in London, 1995 to 2006: case–control study and recommendations. *Eurosurveillance*. 2011 Mar 31;16(13):19830.
402. Littleton J, Park J, Thornley C, Anderson A, Lawrence J. Migrants and tuberculosis: analysing epidemiological data with ethnography. *Australian and New Zealand Journal of Public Health*. 2008 Apr 1;32(2):142–9.
403. Reitmanova S, Gustafson DL. Exploring the Mutual Constitution of Racializing and Medicalizing Discourses of Immigrant Tuberculosis in the Canadian Press. *Qualitative Health Research*. 2012 Jul;22(7):911–20.
404. Pareek M, Greenaway C, Noori T, Munoz J, Zenner D. The impact of migration on tuberculosis epidemiology and control in high-income countries: a review. *BMC Medicine*. 2016 Dec;14(1).
405. Kunimoto D, Sutherland K, Wooldrage K, Fanning A, Chui L, Manfreda J, et al. Transmission characteristics of tuberculosis in the foreign-born and the Canadian-born populations of Alberta, Canada. *The International Journal of Tuberculosis and Lung Disease*. 2004 Oct 1;8(10):1213–20.
406. Storla DG, Yimer S, Bjune GA. A systematic review of delay in the diagnosis and treatment of tuberculosis. *BMC Public Health*. 2008 Jan 14;8(1):15.
407. Vadwai V, Shetty A, Supply P, Rodrigues C. Evaluation of 24-locus MIRU-VNTR in extrapulmonary specimens: Study from a tertiary centre in Mumbai. *Tuberculosis*. 2012 May;92(3):264–72.
408. BC Coroners Service. Illicit Drug Overdose Deaths in BC [Internet]. 2018 Aug [cited 2018 Sep 18]. Available from: <https://www2.gov.bc.ca/assets/gov/birth-adoption-death-marriage-and-divorce/deaths/coroners-service/statistical/illicit-drug.pdf>
409. Ford C, Yusim K, Ioerger T, Feng S, Chase M, Greene M, et al. Mycobacterium tuberculosis – Heterogeneity Revealed Through Whole Genome Sequencing. *Tuberculosis (Edinb)*. 2012 May;92(3):194–201.
410. Lönnroth K, Mor Z, Erkens C, Bruchfeld J, Nathavitharana RR, van der Werf MJ, et al. Tuberculosis in migrants in low-incidence countries: epidemiology and intervention entry points. *The International Journal of Tuberculosis and Lung Disease*. 2017 Jun 1;21(6):624–36.

411. Government of Canada. Annual Report to Parliament on Immigration, 2016 [Internet]. 2016 [cited 2018 Jun 4]. Available from: <https://www.canada.ca/en/immigration-refugees-citizenship/corporate/publications-manuals/annual-report-parliament-immigration-2016.html#abimm>
412. Gushulak BD, Pottie K, Roberts JH, Torres S, DesMeules M. Migration and health in Canada: health in the global village. *Canadian Medical Association Journal*. 2011 Sep 6;183(12):E952–8.
413. BC Centre for Disease Control [creator]. (2015): BC Provincial TB Registry (BCCDC-iPHIS). Population Data BC [publisher]. Data Extract. BCCDC (2014) [Internet]. Available from: <http://www.popdata.bc.ca/data>
414. British Columbia Ministry of Health [creator]. (2015): Consolidation File (MSP Registration & Premium Billing). V2. Population Data BC [publisher]. Data Extract. MOH (2014) [Internet]. Available from: <http://www.popdata.bc.ca/data>
415. Citizenship and Immigration Canada [creator]. (2014): CIC Permanent Residents File. Population Data BC [publisher]. Data Extract. CIC (2015) [Internet]. Available from: <http://www.popdata.bc.ca/data>
416. Ronald LA, Campbell JR, Balshaw RF, Romanowski K, Roth DZ, Marra F, et al. Demographic predictors of active tuberculosis in people migrating to British Columbia, Canada: a retrospective cohort study. *CMAJ*. 2018 Feb 26;190(8):E209–16.
417. Khan K, Hirji MM, Miniota J, Hu W, Wang J, Gardam M, et al. Domestic impact of tuberculosis screening among new immigrants to Ontario, Canada. *Canadian Medical Association Journal*. 2015 Nov 3;187(16):E473–81.
418. Canada, Health Canada. Health Canada’s strategy against tuberculosis for First Nations on-reserve. 2012.
419. Jones M, Ross B, Cloth A, Heller L. Interventions to reach underscreened populations: a narrative review for planning cancer screening initiatives. *Int J Public Health*. 2015 May 1;60(4):437–47.
420. Morris ZS, Wooding S, Grant J. The answer is 17 years, what is the question: understanding time lags in translational research. *Journal of the Royal Society of Medicine*. 2011 Dec;104(12):510–20.

Appendices

Appendix A: Online Pre- and Post-Meeting Survey Questions

Appendix Table A-1. Online survey questions administered pre-meeting.

Questions

A. Basic Information

1. Which tasks related to tuberculosis are generally part of your job? (*Select all that apply*)
 - Supervising daily patient medication doses
 - Collecting contact information directly from patients
 - Writing notes on each patient encounter
 - Entering information into Panorama
 - Case management
 - Outbreak investigation
 - General program oversight
 - None of the above
2. What are your working hours?
 - Full time
 - Part time
 - Other (please specify)
3. Approximately how long have you worked in the area of TB?
 - < 1 year
 - 1–5 years
 - > 5 years
4. Do your daily activities include anything other than TB?
 - Yes
 - No
5. What proportion of your average week do you spend on TB-related work?
 - 0–19%
 - 20–39%
 - 40–59%
 - 60–79%
 - 80–100%

B. TB Genotyping Knowledge

6. Prior to this study had you heard of TB genotyping (*i.e. MIRU-VNTR/spoligotyping/RFLP*)?
 - Yes
 - No
7. Which genotyping methods have you heard of? (*Select all that apply*)
 - RFLP/fingerprinting
 - MIRU-VNTR
 - spoligotyping

Continued on next page

Appendix Table A-1 *Continued from previous page*

Questions

8. Through what means have you been exposed to information about genotyping? (*Select all that apply*)
 - daily work
 - presentations
 - conferences
 - journal articles
 - co-workers
 - reports
 - other
9. Have you ever received formal training (e.g. course, workshop, user guide) in the use of genotyping data for investigations? (*e.g. confirm/refute transmission, guide contact tracing*)
 - Yes
 - No
10. Prior to this study were you aware that genotyping data (MIRU-VNTR) was available for your TB patients?
 - Yes
 - No

C. Current Process

11. Prior to this study, have you used MIRU-VNTR genotyping data in your investigations? (*e.g. confirm/refute transmission, guide contact tracing*)
 - Yes
 - No
 12. Prior to this study, what have you used MIRU-VNTR genotyping data for? (*Select all that apply*)
 - Confirm clusters and links between cases
 - Refute clusters and links between cases
 - Identify unknown links between cases
 - Justify extending contact investigation
 - Investigate potential false positive TB diagnosis
 - Don't know
 - Other (please specify)
 13. Prior to this study, how often do you use MIRU-VNTR genotyping data in your case management or outbreak investigation?
 - Never
 - For few cases
 - For about half of cases
 - For many cases
 - For every case
 14. How confident are you in using MIRU-VNTR genotyping data in your investigations? (*e.g. confirm/refute transmission, guide contact tracing*)
 - Novice – not at all confident
 - Average – somewhat confident
 - Expert – completely confident
-

Appendix Table A-2. Online survey questions administered post-meeting.

Questions

A. Future Processes

1. How often would you like to use MIRU-VNTR genotyping data in your TB case management or outbreak investigations?
 - Never
 - For few cases
 - For about half of cases
 - For many cases
 - For every case
 2. After completion of this study do you feel more confident using MIRU-VNTR genotyping data in your investigations?
 - Not at all
 - Somewhat
 - Considerably
 3. How often would you like to use whole genome sequencing (WGS) data in your case management or outbreak investigation?
 - Never
 - For few cases
 - For about half of cases
 - For many cases
 - For every case
 4. After completion of this study do you feel more confident using whole genome sequencing (WGS) data in your investigations?
 - Not at all
 - Somewhat
 - Considerably
 5. Would you like to receive training in interpretation of TB genotyping and genome sequencing?
 - Yes
 - Maybe
 - No
 6. What format would you prefer for training? (*Select all that apply*)
 - In-person workshops
 - User guides
 - Online videos
 - Other
 7. Reason you do not want training?
 - Not part of my job
 - Not valuable
 - No time
 - Other
 8. Do you have any comments about the survey or the study that you would like to make? We value all input.
-

Appendix Table A-3. Interview questions used to guide discussion.

Category	Questions
Impact	<ul style="list-style-type: none"> • Overall, do you think that genotyping and/or genome sequencing data improved your understanding of transmission dynamics in Yukon? • Do you think that you would have changed your approach to any investigations if you had genotyping and/or genome sequencing data at the time of investigation? • What proportion of cases do you feel that genotyping data and/or genome sequencing data would be helpful? • Overall, do you think that there is added value in genotyping and/or genome sequencing Yukon’s TB isolates?
Future Processes	<ul style="list-style-type: none"> • How should the BCCDC Public Health Laboratory, clinicians, and contact investigation teams communicate genotyping and/or genome sequencing information to YCDC? <ul style="list-style-type: none"> ◦ <i>Data</i>: raw data or analyzed/interpreted data ◦ <i>Format</i>: case-level reports, regular summary reports, regular teleconferences, phone calls as needed • After reviewing the genotype cluster descriptions for Yukon’s MIRU-VNTR clusters we’d like to get your feedback. <ul style="list-style-type: none"> ◦ Did you find the cluster descriptions useful? (<i>Useful / Not very useful / Useless</i>) ◦ What other information would you like to see in the cluster descriptions? ◦ Other feedback?

Appendix B: Presentations

B.1 Oral presentations

1. Whole Genome Sequencing as a Tool to Quantify Local Tuberculosis Transmission in British Columbia, Canada. *39th Annual Congress of European Society of Mycobacteriology*. Dresden, GERMANY, 3 July 2018.
2. Knowledge Translation: Molecular Epidemiology of Tuberculosis in British Columbia. *BC Tuberculosis Clinical Leadership Meeting*. Vancouver, CANADA. 28 May 2018.
3. Whole Genome Sequencing of *Mycobacterium tuberculosis* Identifies Transmission to Pediatric Patients in British Columbia Canada, 2005–2014. *22nd Annual Conference of the Union—North America Region*. Chicago, IL, USA. 2 March 2018.
4. Molecular Epidemiology of Tuberculosis in British Columbia. *Faculty of Medicine Trainees Roundtable*. Vancouver, CANADA. 26 October 2017.
5. A Comparison of Knowledge Gained Through Universal Tuberculosis Genotyping Over the Previous On-Request Program. *BCCDC Research Week Symposium*. Vancouver, CANADA. 26 October 2017.
6. Genotyping and Whole Genome Sequencing to Identify Tuberculosis Transmission Related to Pediatric Patients in British Columbia, Canada, 2005–2014. *48th Union World Conference on Lung Health*. Guadalajara, MEXICO. 12 October 2017.
7. Findings from Whole Genome Sequencing of Tuberculosis in a Geographically Large Canadian Province with a Diverse Population. *38th Congress of European Society of Mycobacteriology*. Šibenik, CROATIA. 28 July 2017.
8. Whole Genome Sequencing for Diagnosis and Surveillance of Tuberculosis in British Columbia. *ASM Microbe 2017*. New Orleans, LA, USA, 5 June 2017.
9. Molecular Epidemiology of Tuberculosis in British Columbia—a 10-year Retrospective Study. *British Columbia Centre for Disease Control Grand Rounds*, Vancouver, BC, CANADA, 2 May 2017.
10. Molecular Epidemiology of Tuberculosis in British Columbia and Yukon—A 10-year Retrospective Study, *Whitehorse General Hospital Grand Rounds*. Whitehorse, YT, CANADA. 27 September 2016.
11. Tuberculosis—Molecular Epidemiology. *TB Molecular Epidemiology Stakeholders' Meeting*. Vancouver, BC, CANADA. 18 April 2016.

B.2 Poster Presentations

1. Whole Genome Sequencing Provides New Insights to Tuberculosis Transmission Over a 10-year Period in British Columbia, Canada. *AMMI Canada—CACMID Annual Conference*. Vancouver, BC, CANADA. 3 May 2018.
2. Whole Genome Sequencing of *Mycobacterium tuberculosis* Identifies Transmission to Pediatric Patients in British Columbia Canada, 2005–2014. *22nd Annual Conference of The Union - North America Region*. Chicago, IL, USA. 3 March 2018.
3. Tuberculosis Transmission in British Columbia, Canada: Insights from Whole Genome Sequencing. *Applied Bioinformatics and Public Health Microbiology*. Hinxton, UK. 18 May 2017.
4. Molecular Epidemiology of Tuberculosis in British Columbia, Canada – A 10-year Retrospective Study. *21st Annual Conference of The Union - North America Region*. Vancouver, BC, CANADA. 24 February 2017.
5. Improved Understanding of Tuberculosis Transmission in British Columbia Using Whole Genome Sequencing. *Faculty of Medicine, UBC—Genes and Environment Workshop*. Vancouver, BC, CANADA. 26 September 2016.
6. Improved Understanding of Tuberculosis Transmission in British Columbia Using Whole Genome Sequencing. *37th Congress of European Society of Mycobacteriology*. Catania, ITALY. 4-5 July 2016.
7. First Steps to Understanding Tuberculosis Transmission in British Columbia: Insights from 24-locus MIRU-VNTR Genotyping. *20th Annual Conference of The Union—North America Region*. Denver, CO, USA. 26 February 2016.
8. Progress Update—A Retrospective Study Utilizing Whole Genome Sequencing to Understand Tuberculosis Transmission in British Columbia. *Applied Bioinformatics and Public Health Microbiology*. Cambridge, UK. 6 May 2015.
9. Progress Update—A Retrospective Study Utilizing Whole Genome Sequencing to Understand Tuberculosis Transmission in British Columbia. *19th Annual Conference of The Union—North America Region*. Vancouver, BC, CANADA. 26 February 2015.

B.3 Media

1. The biomedical science behind tuberculosis treatments [Internet]. Institute of Biomedical Science. [cited 2018 Aug 8]. Available from: <https://www.ibms.org/resources/news/the-biomedical-science-behind-tuberculosis-treatments/>
2. The Poison Terminator – A killer bacteria that won't go away [Internet]. [cited 2018 Aug 8]. Available from: <https://radiopublic.com/the-poison-terminator-8QrMnr/ep/s1!28c45>

Appendix C: Genotyping Summary Report

Copy of: Tuberculosis Genotyping in British Columbia
10-year Retrospective Study Report

Tuberculosis Genotyping in British Columbia

10-year Retrospective Study Report

Prepared by: Jennifer L. Guthrie, PhD Candidate

Email: jennifer.guthrie@alumni.ubc.ca

School of Population and Public Health

University of British Columbia

June 2018



Tuberculosis Genotyping in British Columbia

10-year Retrospective Study Report

Summary

In 2012, a project was initiated by the British Columbia Centre for Disease Control (BCCDC) to retrospectively genotype the first *Mycobacterium tuberculosis* (*Mtb*) isolate from each patient with a culture confirmed diagnosis of tuberculosis (TB) using 24-locus Mycobacterial Interspersed Repetitive Unit - Variable Number Tandem Repeat (MIRU-VNTR). This report describes the resulting cluster analyses and includes the geographical and temporal distribution of large (≥ 10 persons) genotype clusters for *Mtb* isolated from specimens received at the BCCDC Public Health Laboratory (BCCDC PHL) from 2005 through 2014.

Overall, MIRU-VNTR genotyping grouped 2,290 isolates into 189 clusters (2–70 isolates/cluster) with an overall clustering rate of 42.4% and an estimated endemic transmission rate of 34.1% (“*n-1*” method).¹ Large clusters (≥ 10 persons) occurred more frequently within the *Mtb* Euro-American lineage and included mainly Canadian-born persons (87.1%–100.0%). For full details of the *Mtb* molecular epidemiology in British Columbia see Guthrie et al. (2017).²

Key Facts



No. Isolates Genotyped
2,290

No. Distinct Genotypes
1,508

Percentage Clustered
42%

No. Clusters
189

Cluster Size Range
2–70

Canadian-born Clustered
77%

Foreign-born Clustered
30%

Urban Clustered
39%

Rural Clustered
74%



Table of Contents

Summary.....	1
Introduction to Genotyping.....	3
What is 24-locus MIRU-VNTR?.....	4
How is Genotyping Used?.....	4
Limitations.....	5
Data and Analysis.....	6
TB Genotyping in British Columbia Infographic.....	9
MIRU-VNTR Cluster Summaries.....	10
References.....	27
Appendix I: 24-LOCUS MIRU-VNTR PATTERNS OF LARGE CLUSTERS.....	29
Appendix II: MIRU-VNTR ALIASES.....	30



Introduction to Genotyping

Mycobacterium tuberculosis (*Mtb*) genotyping uses DNA based techniques to target specific segments of the genome allowing for the differentiation of *Mtb* strains. Genotyping has a number of public health and research applications, which will be discussed in a later section.

24-locus Mycobacterial Interspersed Repetitive Unit - Variable Number Tandem Repeat (MIRU-VNTR) genotyping has become the standard tool for molecular typing of *Mtb* for many TB programs worldwide. As a rapid technique resulting in a portable digital signature, MIRU-VNTR has replaced genotyping by restriction fragment length polymorphism (RFLP) in most laboratories. A similarly rapid method known as spoligotyping is often used in molecular studies; however, its low resolution makes it unsuitable for inferring transmission. Genomics, which utilizes the entire genome sequence, is the most recent method available and has the highest discriminatory power; however, sequencing technology and analyses have not been fully standardized for routine use and at this time is most often used for research purposes or specific outbreak investigations.

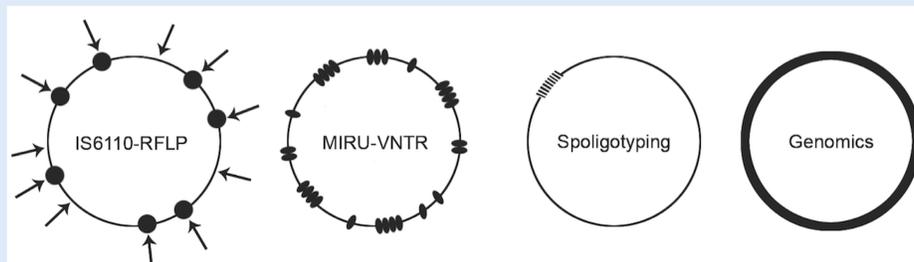


Figure 1. Common Molecular Methods for Genotyping *Mycobacterium tuberculosis*. This simple schematic, not to scale, compares four common methods used in TB molecular epidemiology, and the markings provide an appreciation of the targeted regions for analysis. In contrast, genomics which uses whole genome sequencing interrogates the entire genome, with single nucleotide polymorphisms revealing the relationship between isolates.



What is 24-locus MIRU-VNTR?

The BCCDC PHL uses a standard method³ of 24-locus MIRU-VNTR for routine genotyping. MIRUs (Mycobacterium Interspersed Repetitive Units) represent repeated DNA sequences 40 to 110 base pairs long which are found in a number of locations around the *Mtb* genome.⁴ MIRU-VNTR genotyping is performed by PCR amplification of each MIRU locus using primers specific for the flanking region. Following capillary electrophoresis, the size of each amplicon is determined, and calculations are performed based on the known length of the repeat unit at each locus. The number of repeats at each of the 24 loci are combined to generate a digital signature that can be used to determine the phylogenetic structure and epidemiological links between strains.

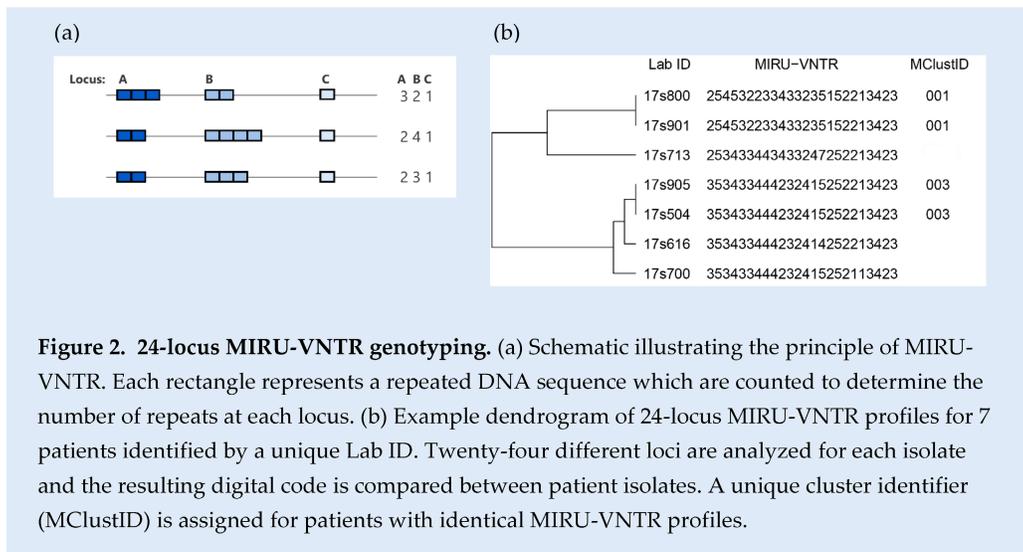


Figure 2. 24-locus MIRU-VNTR genotyping. (a) Schematic illustrating the principle of MIRU-VNTR. Each rectangle represents a repeated DNA sequence which are counted to determine the number of repeats at each locus. (b) Example dendrogram of 24-locus MIRU-VNTR profiles for 7 patients identified by a unique Lab ID. Twenty-four different loci are analyzed for each isolate and the resulting digital code is compared between patient isolates. A unique cluster identifier (MClustID) is assigned for patients with identical MIRU-VNTR profiles.

How is Genotyping Used?

As previously stated, genotyping data has numerous public health and research applications. When combined with epidemiological information *Mtb* genotyping can be a very useful tool. Genotyping results have been used to detect specimen mix-up/laboratory cross-contamination events, identify outbreaks, confirm/refute suspected transmission and differentiate between reinfection and reactivation of tuberculosis. Studies have shown that the routine use of genotyping data enhances contact investigations, leads to more effective use of resources, and can uncover previously unrecognized sources and sites of transmission.^{5,6} Furthermore, genotyping data can be used to monitor clusters over time, evaluate program performance, and understand *Mtb* population dynamics in a particular region or setting.



Limitations

It should be noted that as with any biological test there are limitations to its interpretation. In the case of *Mtb* genotyping, the first limitation is a technical one. Bacterial isolation is required for DNA extraction and genotyping. Consequently, clinically diagnosed cases without culture confirmation (~20% of TB diagnoses in BC) are not able to be genotyped, and therefore their strain type cannot be matched to other cases and will not contribute to the genotyping database. The second issue involves the testing methodology. Standard PCR primers have been designed based on the most common DNA sequences found across *Mtb* strains used during method development but cannot capture all possible sequences that may exist globally and mismatched sequences may result in poor or failed amplification. Moreover, some strains may have a large number of repeats for a particular locus (e.g. MIRU-4052) causing the amplicon size to exceed the upper limit of the genetic analyzer instrument. Genomic rearrangements due to mobile genetic elements may also prevent amplification of some loci (e.g. MIRU-2163 and MIRU-2165). The result of these technical issues is an incomplete MIRU-VNTR pattern which impedes interpretation and, in most cases, does not allow for cluster assignment.

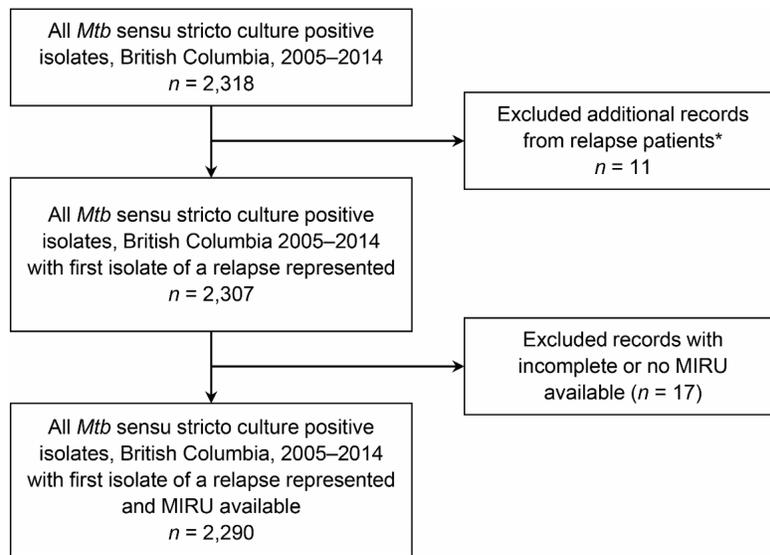
Further limitations involve epidemiological interpretation of genotype information. First, directionality of transmission cannot be determined by genotype data alone, and clustering only indicates that patient strains are genotypically related. Secondly, genotypic clustering does not necessarily mean transmission has occurred between the patients in question, especially in particular patient groups. Whole genome sequencing (WGS) of large MIRU-VNTR clusters elsewhere in Canada, belonging to the Indo-Oceanic lineage revealed that these clusters were not representative of local transmission but rather a common country or region of origin.⁷ In contrast, WGS of large MIRU-VNTR clusters comprised largely of Canadian-born individuals in both BC and Ontario have demonstrated that these clusters represent ongoing local transmission of TB.⁸⁻¹⁰ Ultimately, genotypic clustering should always be interpreted within the context of epidemiological information.



Data and Analysis

The study population included all persons with culture-confirmed TB residing in BC whose first specimen with *Mtb* isolated was received by the BCCDC PHL from 2005 through 2014 ($n = 2,318$). *Mycobacterium africanum*, *Mycobacterium bovis*, and *Mycobacterium bovis* bacilli Calmette-Guérin (BCG) were excluded. For individuals with a recurrence during the study period, data from their first episode only was used if isolates from their first and second episode had matching MIRU-VNTR patterns ($n = 11$), and data from both episodes where MIRU-VNTR indicated reinfection ($n=2$).

Isolates lacking an amplicon peak at any locus were repeated with newly extracted DNA, and where there remained no peak at a single locus – excluding MIRU-VNTR loci 2163 and 2165, which are treated as absent when there is no amplification¹¹ – the locus was coded as missing data and included in the analyses ($n = 93$). Of the 2,307 culture-positive isolates meeting study criteria (Figure 3), 17 isolates had incomplete MIRU-VNTR patterns or were unavailable for genotyping – leaving a total of 2,290 (99.2%) isolates which were successfully genotyped by 24-locus MIRU-VNTR using standard methods.³



*First episode for a relapse patient was maintained in the study; relapse was defined as a subsequent episode with a genotype ≤ 1 MIRU loci different to the initial episode.

Figure 3. Analytic Sample. Selection of the analytic sample to examine the molecular epidemiology of tuberculosis in British Columbia, 2005–2014.

Tuberculosis Genotyping in British Columbia



Isolates with an identical 24-locus MIRU-VNTR pattern were assigned an “MClust” number (a unique cluster ID) representing a unique genotypic cluster (≥ 2 individuals). Large clusters (≥ 10 cases) were analyzed to determine the predominant birthplace (Canada or Outside Canada), and were assigned as Canada where $>50\%$ of persons in the cluster were born in Canada, otherwise the predominant birthplace was classified as Outside Canada. Cluster composition for each cluster in the study was categorized as: (i) exclusively Canadian-born, (ii) exclusively foreign-born, (iii) mixed Canadian- and foreign-born, or (iv) unknown. “Unknown” was defined as a cluster in which 1 or more individuals’ birthplace was not known and the remaining clustered individuals were uniformly born in Canada or Outside Canada.

Individual-level clinical and demographic data were extracted from BCCDC’s Integrated Provincial Health Information System (iPHIS). Community type was determined using the population density of the geographic service area in which each patient resided – urban ($>40,000$), or rural ($\leq 40,000$).

Mtb can be classified into seven major phylogeographic lineages reflective of the coevolution of tuberculosis and humans, and linked to ancient human migration patterns.^{12,13} As a result, lineage information provides additional epidemiological information which contributes to the overall understanding of the *Mtb* population dynamics in a setting, and may contribute to case investigations – acting as an alert where lineage does not match what is expected based on a patient’s demographics and travel history. Here, major lineage was predicted for each isolate based on MIRU-VNTR using TB-Insight’s CBN method.¹⁴ Phylogenetic relationships within each major lineage were visualized using a minimum-spanning tree (MST) in PHYLOViZ 2.0¹⁵ and were coloured by birthplace (Figure 4).

Tuberculosis Genotyping in British Columbia

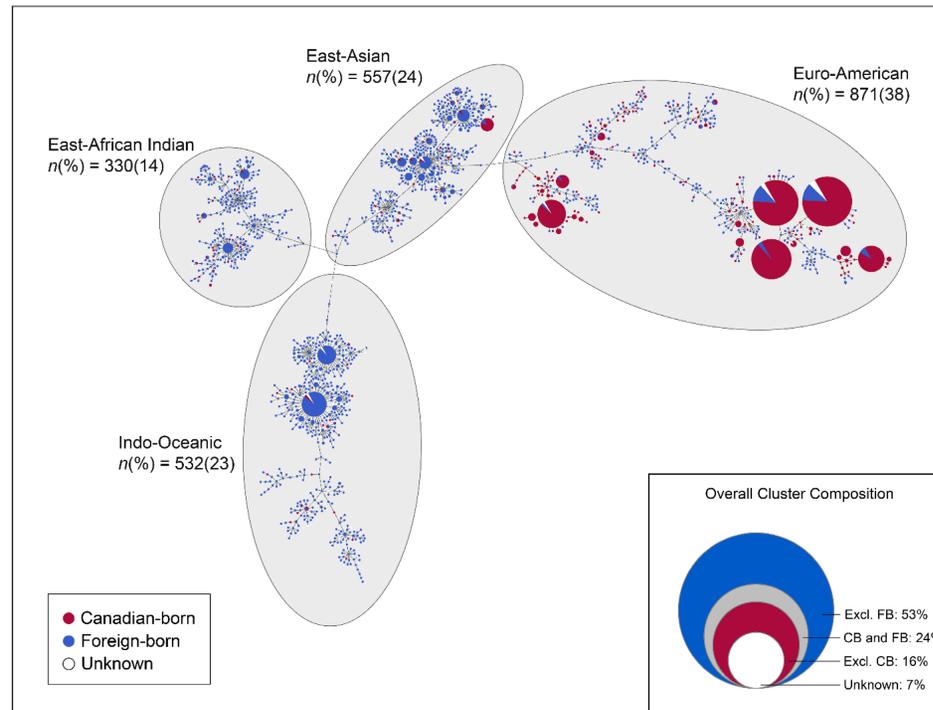


Figure 4. Minimum spanning tree analysis of 24-locus MIRU-VNTR genotyping for *Mycobacterium tuberculosis* isolates, British Columbia (2005–2014). The size of each circle is proportional to the number of isolates. Classification of strains by birthplace is visualized by color coding. The inset demonstrates overall cluster composition with respect to birthplace; relative frequency of clusters that were exclusively Canadian-born (Excl.CB), exclusively foreign-born (Excl. FB), Canadian- and foreign-born (CB and FB), or where there were cases in a cluster with only CB or FB identified in addition to ≥ 1 case of unknown birthplace. *Percentages have been rounded and may not total to 100%.



TUBERCULOSIS GENOTYPING IN BRITISH COLUMBIA, 2005-2014

2012 **GOAL**

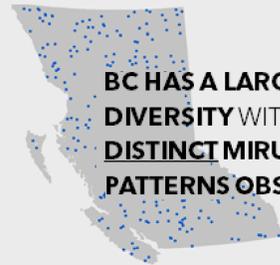


TB INCIDENCE IN **10-YEARS**

GENOTYPING SUPPORTS THIS GOAL BY **INFORMING** CONTACT INVESTIGATIONS OF TB STRAIN **MATCHES**

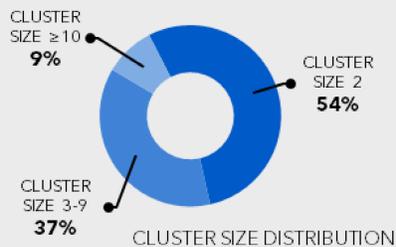


OF **2,290** ISOLATES GENOTYPED **42%** CLUSTERED, FOR A RECENT TRANSMISSION ESTIMATE OF **34%**



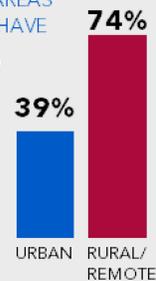
BC HAS A LARGE STRAIN DIVERSITY WITH **1,508** DISTINCT MIRU-VNTR PATTERNS OBSERVED

BC's LARGEST CLUSTER HAS **70** PERSONS



PATIENTS RESIDING IN **RURAL** and **REMOTE** AREAS ARE MORE LIKELY TO HAVE

CLUSTERED STRAINS

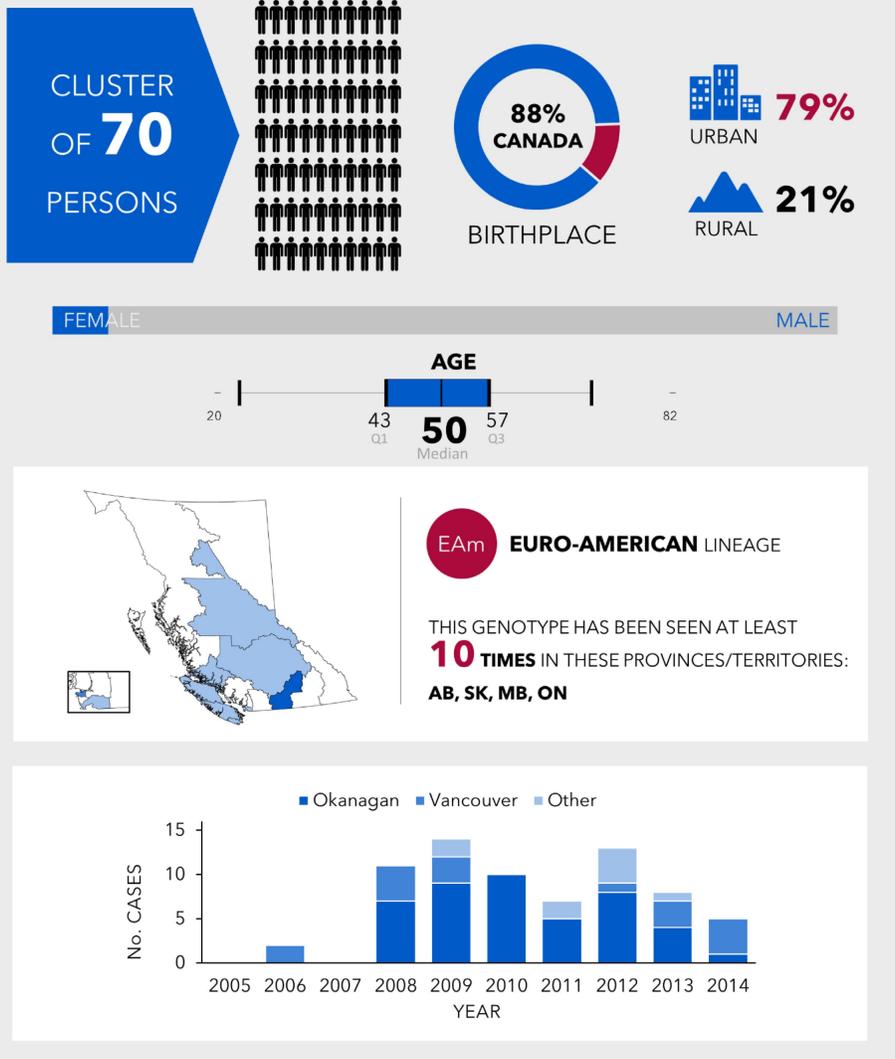




Large MIRU-VNTR Cluster Summaries



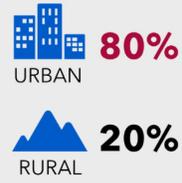
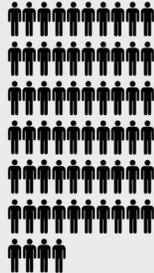
MIRU-VNTR CLUSTER SUMMARY MCLUST-002





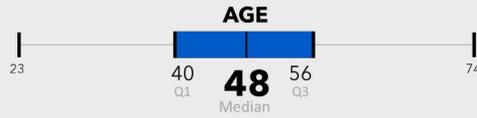
MIRU-VNTR CLUSTER SUMMARY MCLUST-012

CLUSTER
OF **64**
PERSONS



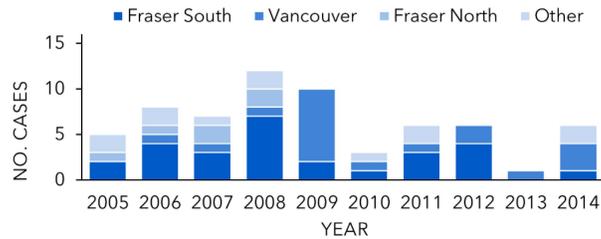
FEMALE

MALE



EAm **EURO-AMERICAN** LINEAGE

THIS GENOTYPE HAS BEEN SEEN AT LEAST
33 TIMES IN THESE PROVINCES/TERRITORIES:
AB, SK, MB, ON, NWT, YT





MIRU-VNTR CLUSTER SUMMARY MCLUST-001

CLUSTER
OF **56**
PERSONS

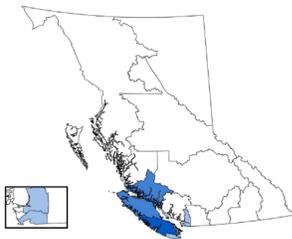
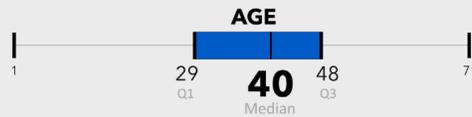


BIRTHPLACE



FEMALE

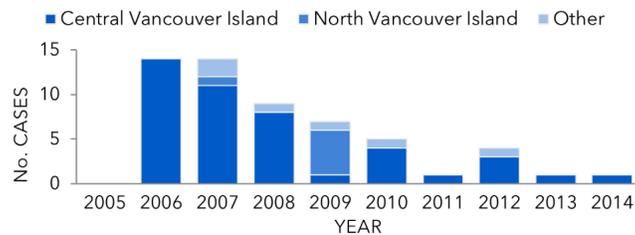
MALE



EAm EURO-AMERICAN LINEAGE

THIS GENOTYPE HAS BEEN SEEN AT LEAST
1 TIMES IN THESE PROVINCES/TERRITORIES:

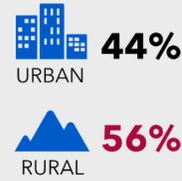
AB





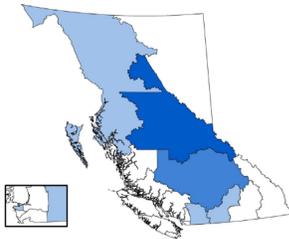
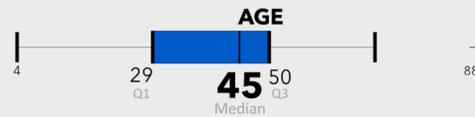
MIRU-VNTR CLUSTER SUMMARY MCLUST-003

CLUSTER
OF **39**
PERSONS



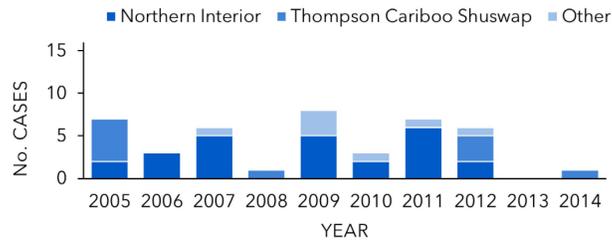
FEMALE

MALE



EAm EURO-AMERICAN LINEAGE

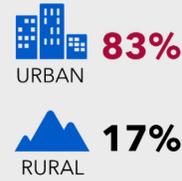
THIS GENOTYPE HAS BEEN SEEN AT LEAST
4 TIMES IN THESE PROVINCES/TERRITORIES:
AB, MB





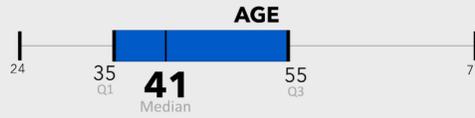
MIRU-VNTR CLUSTER SUMMARY MCLUST-008

CLUSTER
OF **36**
PERSONS



FEMALE

MALE



EAm EURO-AMERICAN LINEAGE

THIS GENOTYPE HAS BEEN SEEN AT LEAST
13 TIMES IN THESE PROVINCES/TERRITORIES:
AB, MB, YT



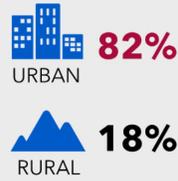


MIRU-VNTR CLUSTER SUMMARY MCLUST-035

CLUSTER
OF **17**
PERSONS

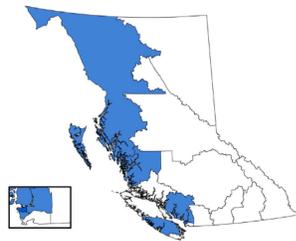
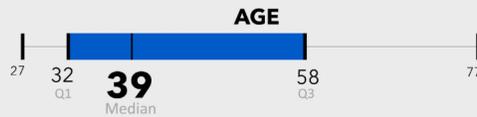


BIRTHPLACE



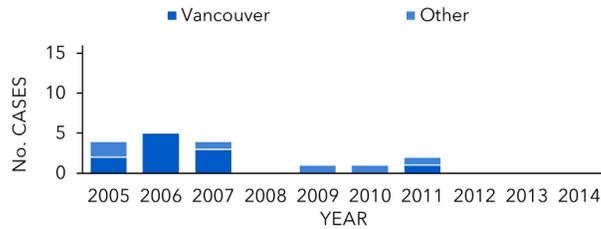
FEMALE

MALE



EA **EAST-ASIAN** LINEAGE

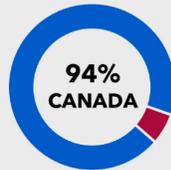
THIS GENOTYPE HAS BEEN SEEN AT LEAST
2 TIMES IN THESE PROVINCES/TERRITORIES:
AB, ON





MIRU-VNTR CLUSTER SUMMARY MCLUST-052

CLUSTER
OF **17**
PERSONS

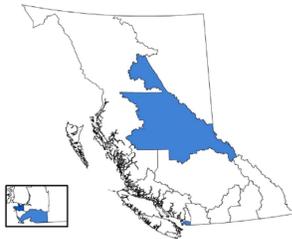
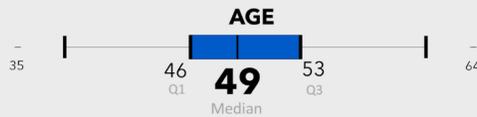


BIRTHPLACE



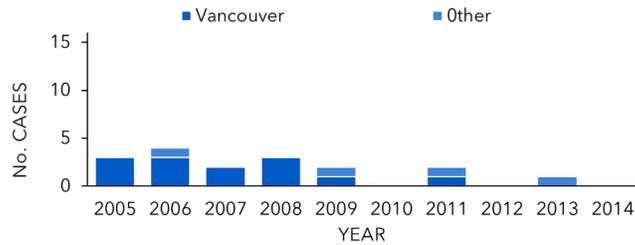
FEMALE

MALE



EAm EURO-AMERICAN LINEAGE

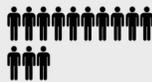
THIS GENOTYPE HAS BEEN SEEN AT LEAST
8 TIMES IN THESE PROVINCES/TERRITORIES:
AB, SK, MB, ON, YT





MIRU-VNTR CLUSTER SUMMARY MCLUST-134

CLUSTER
OF **13**
PERSONS



BIRTHPLACE



URBAN

15%

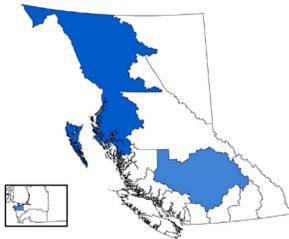
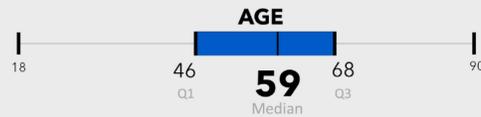


RURAL

85%

FEMALE

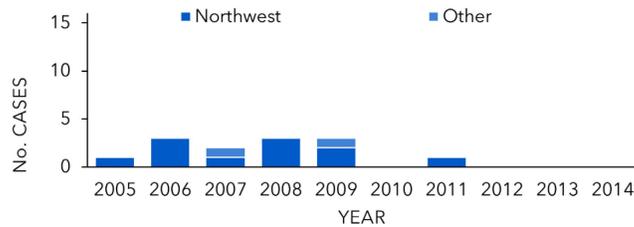
MALE



EURO-AMERICAN LINEAGE

THIS GENOTYPE HAS BEEN SEEN AT LEAST
1 TIME IN THESE PROVINCES/TERRITORIES:

AB





MIRU-VNTR CLUSTER SUMMARY MCLUST-055

CLUSTER
OF **10**
PERSONS

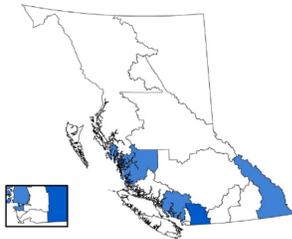
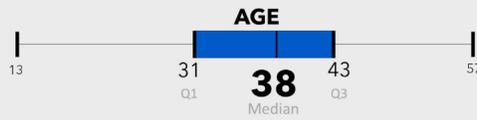


BIRTHPLACE



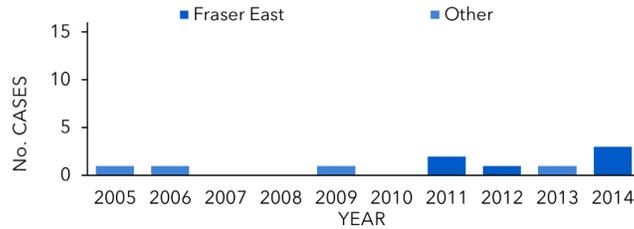
FEMALE

MALE



EAm EURO-AMERICAN LINEAGE

THIS GENOTYPE HAS BEEN SEEN AT LEAST
4 TIMES IN THESE PROVINCES/TERRITORIES:
AB, QC



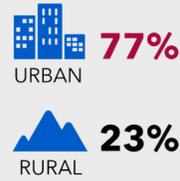


MIRU-VNTR CLUSTER SUMMARY MCLUST-011

CLUSTER
OF **34**
PERSONS

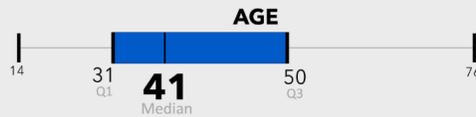


BIRTHPLACE



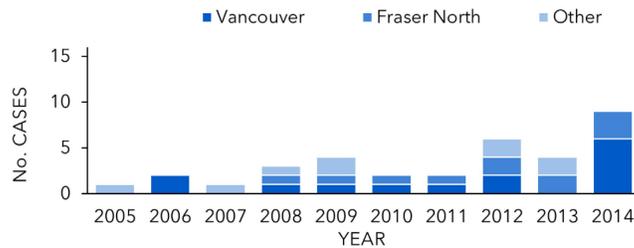
FEMALE

MALE



IO **INDO-OCEANIC** LINEAGE

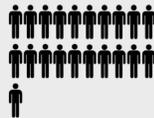
THIS GENOTYPE HAS BEEN SEEN AT LEAST
146 TIMES IN THESE PROVINCES/TERRITORIES:
AB, SK, MB, ON, QC



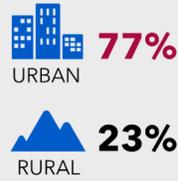


MIRU-VNTR CLUSTER SUMMARY MCLUST-021

CLUSTER
OF **21**
PERSONS

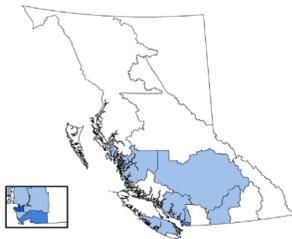
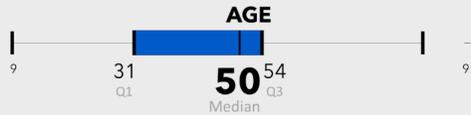


BIRTHPLACE



FEMALE

MALE



IO INDO-OCEANIC LINEAGE

THIS GENOTYPE HAS BEEN SEEN AT LEAST **119** TIMES IN THESE PROVINCES/TERRITORIES:
AB, SK, MB, ON, QC



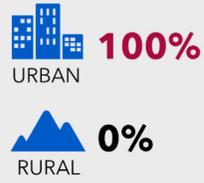


MIRU-VNTR CLUSTER SUMMARY MCLUST-038

CLUSTER
OF **16**
PERSONS



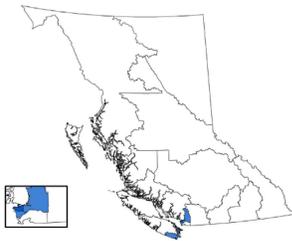
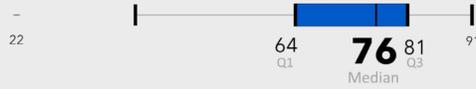
BIRTHPLACE



FEMALE

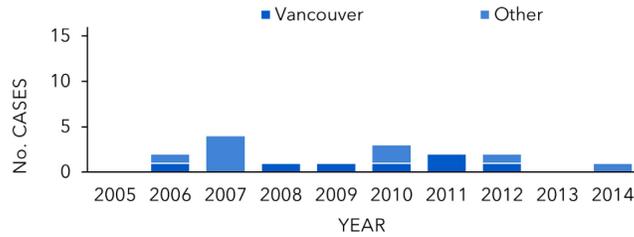
MALE

AGE



EA **EAST-ASIAN** LINEAGE

THIS GENOTYPE HAS BEEN SEEN AT LEAST
48 TIMES IN THESE PROVINCES/TERRITORIES:
AB, MB, ON, QC, NS





MIRU-VNTR CLUSTER SUMMARY MCLUST-187

CLUSTER
OF **16**
PERSONS



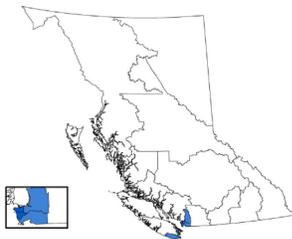
BIRTHPLACE



FEMALE

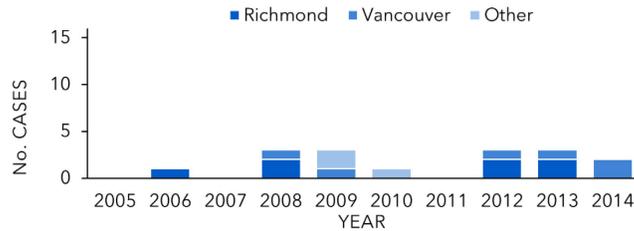
MALE

AGE



EA **EAST-ASIAN** LINEAGE

THIS GENOTYPE HAS BEEN SEEN AT LEAST
65 TIMES IN THESE PROVINCES/TERRITORIES:
AB, SK, MB, ON, QC





MIRU-VNTR CLUSTER SUMMARY MCLUST-149

CLUSTER
OF **13**
PERSONS



BIRTHPLACE



62%

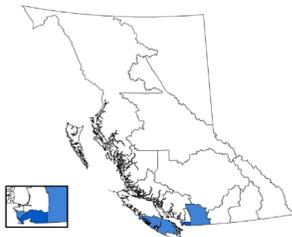
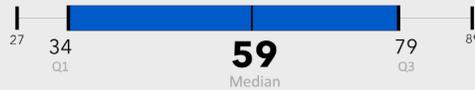


38%

FEMALE

MALE

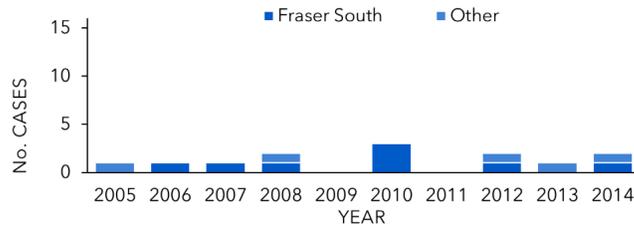
AGE



EAI

EAST-AFRICAN INDIAN LINEAGE

THIS GENOTYPE HAS BEEN SEEN AT LEAST
22 TIMES IN THESE PROVINCES/TERRITORIES:
AB, ON, QC





MIRU-VNTR CLUSTER SUMMARY MCLUST-046

CLUSTER
OF **12**
PERSONS



BIRTHPLACE



92%

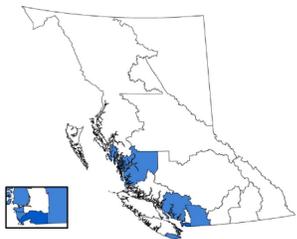


8%

FEMALE

MALE

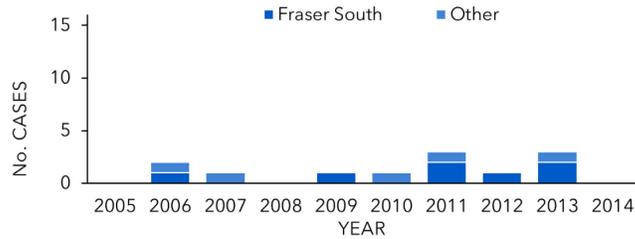
AGE



EAI

EAST-AFRICAN INDIAN LINEAGE

THIS GENOTYPE HAS BEEN SEEN AT LEAST
37 TIMES IN THESE PROVINCES/TERRITORIES:
AB, MB, ON, QC





MIRU-VNTR CLUSTER SUMMARY MCLUST-032

CLUSTER
OF **11**
PERSONS

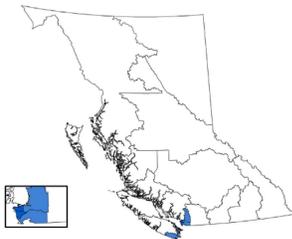
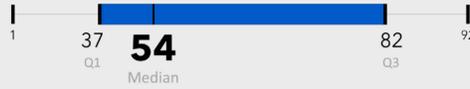


BIRTHPLACE

FEMALE

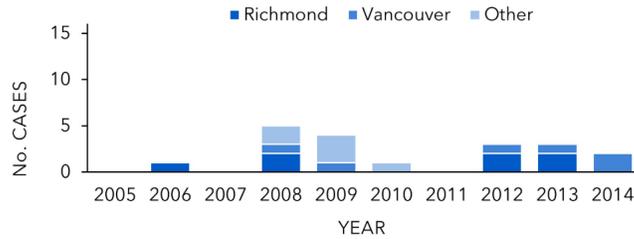
MALE

AGE



EA **EAST-ASIAN** LINEAGE

THIS GENOTYPE HAS BEEN SEEN AT LEAST
20 TIMES IN THESE PROVINCES/TERRITORIES:
AB, ON, QC





References

1. Small PM, Hopewell PC, Singh SP, Paz A, Parsonnet J, Ruston DC, et al. The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods. *N Engl J Med*. 1994 Jun 16;330(24):1703–9.
2. Guthrie JL, Kong C, Roth D, Jorgensen D, Rodrigues M, Hoang L, et al. Molecular Epidemiology of Tuberculosis in British Columbia, Canada: A 10-Year Retrospective Study. *Clin Infect Dis*. 2017 Oct 21;66(6):849–856.
3. Supply P, Allix C, Lesjean S, Cardoso-Oelemann M, Rüsch-Gerdes S, Willery E, et al. Proposal for Standardization of Optimized Mycobacterial Interspersed Repetitive Unit-Variable-Number Tandem Repeat Typing of *Mycobacterium tuberculosis*. *J Clin Microbiol*. 2006 Dec 1;44(12):4498–510.
4. Supply P, Mazars E, Lesjean S, Vincent V, Gicquel B, Locht C. Variable human minisatellite-like regions in the *Mycobacterium tuberculosis* genome. *Mol Microbiol*. 2000 May 1;36(3):762–71.
5. Clark CM, Driver CR, Munsiff SS, Driscoll JR, Kreiswirth BN, Zhao B, et al. Universal Genotyping in Tuberculosis Control Program, New York City, 2001–2003. *Emerg Infect Dis*. 2006 May;12(5):719–24.
6. Lambregts-van Weezenbeek CSB, Sebek MMGG, van Gerven PJHJ, de Vries G, Verver S, Kalisvaart NA, et al. Tuberculosis contact investigation and DNA fingerprint surveillance in The Netherlands: 6 years' experience with nation-wide cluster feedback and cluster monitoring. *Int J Tuberc Lung Dis*. 2003 Dec 1;7(12):S463–70.
7. Jamieson FB, Teatero S, Guthrie JL, Neemuchwala A, Fittipaldi N, Mehaffy C. 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*. 2014 Oct;52(10):3795–8.
8. Cheng JM, Hiscoe L, Pollock SL, Hasselback P, Gardy JL, Parker R. xA clonal outbreak of tuberculosis in a homeless population in the interior of British Columbia, Canada, 2008–2015. *Epidemiol Infect*. 2015 Nov;143(15):3220–3226.
9. Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L, Brodtkin E, et al. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N Engl J Med*. 2011 Feb 24;364(8):730–9.
10. Mehaffy C, Guthrie JL, Alexander DC, Stuart R, Rea E, Jamieson FB. Marked microevolution of a unique *Mycobacterium tuberculosis* strain in 17 years of ongoing transmission in a high risk population. *PLoS One*. 2014;9(11):e112928.

Tuberculosis Genotyping in British Columbia



11. Menéndez MC, Buxton RS, Evans JT, Gascoyne-Binzi D, Barlow REL, Hinds J, et al. Genome analysis shows a common evolutionary origin for the dominant strains of *Mycobacterium tuberculosis* in a UK South Asian community. *Tuberculosis*. 2007 Sep 1;87(5):426–36.
12. Firdessa R, Berg S, Hailu E, Schelling E, Gumi B, Erenso G, et al. Mycobacterial lineages causing pulmonary and extrapulmonary tuberculosis, Ethiopia. *Emerg Infect Dis*. 2013 Mar;19(3):460–3.
13. Gagneux S. Host–pathogen coevolution in human tuberculosis. *Philos Trans R Soc Lond B Biol Sci*. 2012 Mar 19;367(1590):850–9.
14. Shabbeer A, Cowan LS, Ozcaglar C, Rastogi N, Vandenberg SL, Yener B, et al. TB-Lineage: an online tool for classification and analysis of strains of *Mycobacterium tuberculosis* complex. *Infect Genet Evol*. 2012 Jun;12(4):789–97.
15. Francisco AP, Vaz C, Monteiro PT, Melo-Cristino J, Ramirez M, Carriço JA. PHYLOViZ: phylogenetic inference and data visualization for sequence based typing methods. *BMC Bioinformatics*. 2012;13:87.

Tuberculosis Genotyping in British Columbia



Appendix I: 24-LOCUS MIRU-VNTR PATTERNS OF LARGE CLUSTERS

MClustID	MIRU 02	MIRU 04	MIRU 10	MIRU 16	MIRU 20	MIRU 23	MIRU 24	MIRU 26	MIRU 27	MIRU 31	MIRU 39	MIRU 40	424	577	1955	2163	2165	2347	2401	2461	3171	3690	4052	4156	MIRU (InternationalOrder)*
MClust-002	2	2	4	3	2	5	1	5	3	3	2	3	4	4	4	2	3	4	4	2	3	3	7	3	253433443433247252213423
MClust-012	2	2	5	3	2	5	1	5	3	3	2	3	2	3	3	3	3	4	4	2	3	3	5	3	253533233433335252213423
MClust-001	1	2	5	3	2	5	1	5	3	2	2	4	2	3	3	2	3	4	4	2	3	3	5	3	254532233433235152213423
MClust-003	2	3	4	3	2	5	1	5	3	3	2	3	4	4	1	4	4	4	2	2	3	3	5	2	353433444232415252213423
MClust-008	2	2	8	2	2	5	1	1	3	2	2	2	3	4	3	2	4	4	4	3	3	4	8	3	212822344443238252213433
MClust-035	2	2	3	3	2	5	1	6	3	5	3	3	2	4	5	5	4	4	4	2	3	3	8	2	263335244432558253213423
MClust-052	2	3	2	3	2	5	1	5	3	3	2	4	3	4	1	4	4	4	2	2	3	3	9	2	354233344232419252213423
MClust-134	2	2	8	2	2	5	1	1	3	2	2	1	3	4	3	2	4	4	4	2	3	3	8	3	211822344433238252213423
MClust-055	2	2	5	3	1	3	1	5	3	3	2	1	2	3	3	6	3	2	4	2	3	3	7	3	251533233433637232113223
MClust-038	2	2	3	3	2	5	1	7	3	5	3	3	4	4	5	5	4	4	4	2	3	3	8	2	273335444432558253213423
MClust-187	2	2	3	3	2	5	1	7	3	5	3	3	4	4	5	6	4	4	4	2	3	3	8	2	273335444432658253213423
MClust-046	2	2	6	4	2	5	1	7	3	4	2	3	4	2	4	2	4	4	4	2	3	4	7	4	273644424444247252213423
MClust-149	2	2	5	4	2	5	1	7	3	5	3	3	5	2	4	2	4	4	2	2	3	3	8	4	273545524234248253213423
MClust-011	2	5	4	3	2	6	2	2	3	4	3	2	1	4	10	8	4	3	2	6	3	2	7	1	5224341442218A7263223363
MClust-021	2	5	4	3	2	6	2	2	3	4	3	2	1	4	10	9	4	3	2	6	3	2	7	1	5224341442219A7263223363
MClust-032	2	2	2	3	2	5	1	7	3	5	4	3	4	4	5	6	4	4	4	2	3	3	8	2	273235444432658254213423

*International order of loci: MIRU 04, MIRU 26, MIRU 40, MIRU 10, MIRU 16, MIRU 31, 424, 577, 2165, 2401, 3690, 4156, 2163, 1955, 4052, MIRU 02, MIRU 23, MIRU 39, MIRU 20, MIRU 24, MIRU 27, 2347, 2461, 3171



Appendix II: MIRU-VNTR ALIASES

Locus	Alias1	Alias2	12-locus	15-locus	24-locus
154	MIRU 02		x		x
424	Mtub04			x	x
577	ETRC			x	x
580	MIRU 04	ETRD	x	x	x
802	MIRU 40		x	x	x
960	MIRU 10		x	x	x
1644	MIRU 16		x	x	x
1955	Mtub21			x	x
2059	MIRU 20		x		x
2163b	QUB11b			x	x
2165	ETRA			x	x
2347	Mtub29				x
2401	Mtub30			x	x
2461	ETRB				x
2531	MIRU 23		x		x
2687	MIRU 24		x		x
2996	MIRU 26		x	x	x
3007	MIRU 27	QUB5	x		x
3171	Mtub34				x
3192	MIRU 31	ETRE	x	x	x
3690	Mtub39			x	x
4052	QUB26			x	x
4156	QUB4156			x	x
4348	MIRU 39		x		x