

RESEARCH ARTICLE

TBpore cluster: A novel phylogenetic pipeline for tuberculosis transmission studies using nanopore next-generation sequencing data

Sophie Gagnon^{1,2}, Emmanuelle Ametepe^{1,3}, Floriane Point¹, William Cloutier Charette², Arpita Chakravarti², Paul Rivest⁴, Pierre-Marie Akochy⁵, Hafid Soualhine⁶, Zamin Iqbal^{7^{ab}}, Michael B. Hall^{7^{ab}}, Simon Grandjean Lapierre^{1,2^{ab}*}

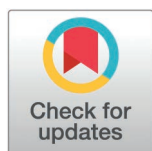
1 Immunopathology Axis, Centre de Recherche du Centre Hospitalier de l'Université de Montréal, Montréal, Québec, Canada, **2** Department of Microbiology, Infectious Diseases and Immunology, Université de Montréal, Montréal, Québec, Canada, **3** École de Santé Publique de l'Université de Montréal, Université de Montréal, Montréal, Québec, Canada, **4** Direction de Santé Publique, Centre Intégré Universitaire de Soins et de Services Sociaux du Centre-Sud-de-l'Île-de-Montréal, Montréal, Québec, Canada, **5** Laboratoire de Santé Publique du Québec, Montréal, Québec, Canada, **6** National Microbiology Laboratory, Winnipeg, Canada, **7** European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridgeshire, United Kingdom

☞ These authors contributed equally to this work.

^{ab} Current address: University of Bath, Bath, United Kingdom.

^{ab} Current address: Department of Microbiology and Immunology, The Peter Doherty Institute for Infection and Immunity, The University of Melbourne, Melbourne, Australia.

* simon.grandjean.lapierre@umontreal.ca



OPEN ACCESS

Citation: Gagnon S, Ametepe E, Point F, Cloutier Charette W, Chakravarti A, Rivest P, et al. (2025) TBpore cluster: A novel phylogenetic pipeline for tuberculosis transmission studies using nanopore next-generation sequencing data. PLoS One 20(6): e0325914. <https://doi.org/10.1371/journal.pone.0325914>

Editor: Herman Tse, Khoo Teck Puat Hospital, SINGAPORE

Received: December 9, 2024

Accepted: May 20, 2025

Published: June 16, 2025

Copyright: © 2025 Gagnon et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data availability statement: Illumina and Nanopore sequences from this study are available in ENA-Biostudies under accession numbers ENA-Biostudies S-BSST1485. TBpore code is accessible from GitHub at <https://github.com/mbhall88/tbpore>.

Abstract

Background

Molecular typing of *Mycobacterium tuberculosis* complex isolates enhances understanding of tuberculosis (TB) transmission dynamics, supporting public health efforts in outbreak investigations. This study aims to validate TBpore, a novel bioinformatic pipeline for clustering TB transmission isolates using Oxford Nanopore Technology (ONT) data and comparing it against conventional Mycobacterial Interspersed Repetitive-Unit Variable Number (MIRU-VNTR) typing and Illumina sequencing.

Methodology/Principal findings

This retrospective case-control study included 58 clinical isolates from two TB outbreaks in Canada, previously characterized by public health investigations and MIRU-VNTR typing. DNA extraction and sequencing were performed on both Illumina and ONT platforms. Illumina data were processed using Clockwork and psdm, while Nanopore data were analyzed with TBpore. SNP distances were used to compare clustering results across methods, with clusters defined by SNP distance thresholds of ≤ 5 and ≤ 12 . Both sequencing methods showed a high degree of concordance in clustering results. All isolates from the *M. africanum* outbreak clustered within the defined SNP thresholds, consistent with MIRU-VNTR and epidemiological data. In

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors declare no financial, personal, or professional interests that could be construed to have influenced the work.

the *M. tuberculosis* outbreak, 20 out of 21 isolates clustered similarly across methods, with one exception. Within outbreak pairwise SNP distances were lower with Nanopore.

Conclusion/Significance

ONT sequencing and the TBpore pipeline offer an accurate alternative to Illumina technology for TB molecular epidemiology. This study suggests potential increased clustering sensitivity with Nanopore technology, warranting further validation on larger datasets with robust epidemiological metadata.

Introduction

Mycobacterium tuberculosis complex molecular typing can improve our understanding of disease transmission dynamics and hence support public health systems in outbreak investigations and contact tracing [1]. *Mycobacterium tuberculosis* (TB) whole genome sequencing (WGS) is progressively replacing previous PCR-based molecular typing methods including MIRU-VNTR due to its increased resolution [2]. WGS is hence already implemented in a systematic and prospective approach in several public health laboratories (e.g., England, Netherlands, New York State) [3,4].

Oxford Nanopore Technology (ONT) next-generation sequencing (NGS) platforms, including the portable MinION sequencer, are increasingly used in TB diagnostics. Hall *et al.* recently compared ONT-based TB sequencing to the benchmark reference Illumina technology for its ability to predict drug resistance and cluster genomically related isolates [5]. This sequencing platform head-to-head comparative analysis did not include clinical isolates for which person-to-person transmission or an epidemiological outbreak context had been confirmed by conventional clinical and epidemiology investigations.

In this study, we report on the clinical validation of TBpore, a new publicly available bioinformatic analysis pipeline enabling molecular clustering of putative transmission isolates. TBpore is designed as a companion tool to the ONT compatible genotypic drug susceptibility testing pipeline Mykrobe enabling simultaneous drug resistance and clustering testing [6]. This is a retrospective case-control molecular epidemiology study to measure the clustering ability and resolution of TBpore compared to conventional MIRU-VNTR and established Illumina NGS-based molecular typing methods. This study includes isolates from domestic TB outbreaks in Canada, which were previously well characterized by routine clinical and public health investigations [7].

Results

In the *M. africanum* outbreak, both WGS technologies and pipelines identified 4 isolates as lineage 5 (L5) and 12 isolates as lineage 6 (L6) using SNP-typing. Four L6 isolates had identical MIRU-type (2 2 4 4 2 2 2 4 3 5 2 2 3 5 4 5 6 2 4 4 3 3 2 4) and among those, respectively, three (MB066247, MB077499, MB086368) and one (MB085690) had a strong and moderate probability of being epidemiologically related

according to field investigations [7]. In the *M. tuberculosis* lineage 4 (L4) outbreak, 21 isolates shared a MIRU-type (2 2 4 3 2 5 1 4 3 3 2 5 2 3 4 5 3 4 4 2 3 4 7 3). All patients were from the same Montreal area and shared one or more risk factors, including alcohol and/or drug use, supervised drug injection sites visits, homelessness or prostitution. In this putative outbreak context, it was not possible to establish direct nominal epidemiological links between all patients.

The TBpore (Nanopore) and psdm (Illumina) derived SNP distances matrices showed a high level of agreement (Fig 1, S1 in S1 File). For the *M. africanum* outbreak, all 4 isolates previously clustered by MIRU-VNTR and conventional epidemiological investigation remained within the established thresholds when analyzed with both sequencing methods. Although one of those four genotypically clustered isolates (MB085690) was not as strongly related to the others according to public health investigation, all three typing approaches suggested its molecular relatedness. Similarly with the *M. tuberculosis* outbreak and control isolates, all pairwise comparison between the 21 isolates with matching MIRU-types fell within the established SNP threshold when analyzed with both methods, except for one isolate pair (MB086444, L00208036) (Fig 2, S1 in S1 File). This pair exceeded the established threshold by one SNP (13 SNP distance) when sequenced with Illumina and analyzed using COMPASS+psdm, but fell within the 12-SNP threshold when sequenced

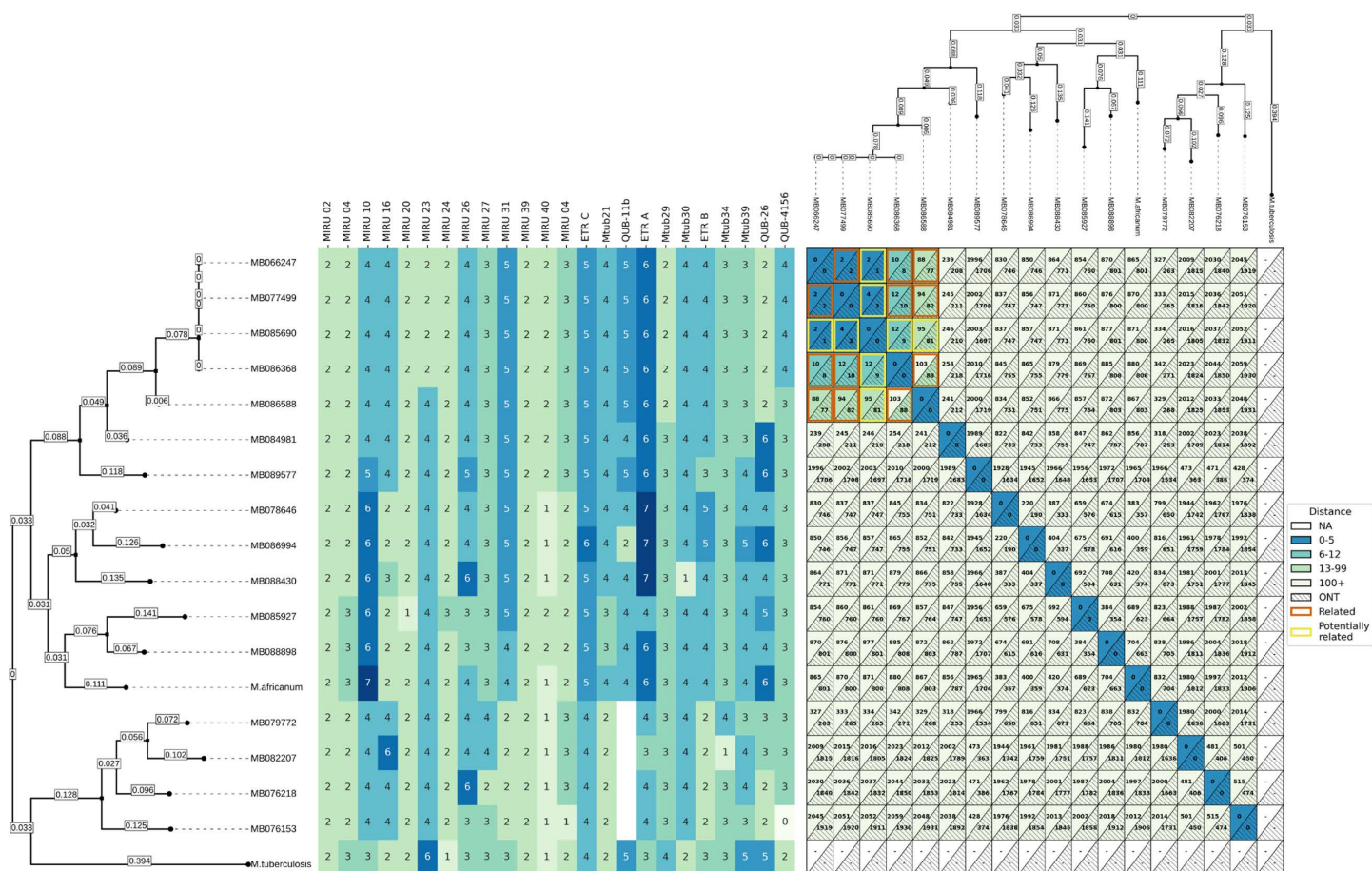


Fig 1. Whole genome sequencing and MIRU-VNTR typing results of *M. africanum* lineage 5 and lineage 6 outbreak. WGS and MIRU-VNTR typing results of *M. africanum* outbreaks, depicting the clustering of lineage 5 (L5) and lineage 6 (L6) isolates. Pairwise SNP distances were calculated using both Illumina (psdm) and Nanopore (TBpore) sequencing methods. Isolates from the same outbreak cluster within the SNP thresholds of ≤ 5 and ≤ 12 transmission clusters, aligning with MIRU-VNTR genotyping and epidemiological data. Reference *M. africanum* (ATCC25420) and *M. tuberculosis* (ATCC27294) are included as controls.

<https://doi.org/10.1371/journal.pone.0325914.g001>

using the Nanopore platform and TBpore cluster pipeline. Except for one isolate pair (MB090711, MB090710), all *M. tuberculosis* control isolates were shown to be genotypically unrelated to each other as well as with isolates from the known person-to-person transmission event (Figs 1 & 2, S1 in S1 File). Those two isolates had distinct MIRU-VNTR profile and had not been identified as a potential domestic transmission event by public health teams but were respectively 7 and 20 SNP apart when sequenced and analysed with Illumina and ONT platforms and pipelines.

In the *M. africanum* outbreak and control isolates dataset, the Illumina SNP distances ranged from 2 to 2059, while Nanopore's ranged from 1 to 1931 (Fig 2). The absolute differences in pairwise SNP distances between Illumina and Nanopore ranged from 0 to 432 SNPs, with an average absolute difference of 127 SNPs. Using a paired permutation statistical test and a significance threshold of 0.05, the pairwise SNP distances were found to be significantly higher with Illumina sequencing and the COMPASS analysis pipeline. Similar observations were made with the *M. tuberculosis* data (see S2 in S1 File).

Discussion

WGS-based molecular typing is increasingly used by public health departments due to the comprehensive insights into TB transmission dynamics it provides as well as its increasing affordability and turnaround time. The diversification and democratization of sequencing technologies and analysis pipelines is ongoing and will help better address the specific needs of different settings. ONT platform acquisition costs are lower than those of Illumina, and ONT protocols and workflows are better suited for smaller sample processing volumes. ONT may hence represent a better strategy for laboratories and public health programs in limited resource settings or those with low TB incidence without other sequencing needs. The impact of those technology-specific relative advantages remains underexplored and merit further evaluation once WGS-based molecular investigations will be performed at scale in reference laboratory and public health workflows including those of resource-limited settings.

This study confirms the ability of ONT sequencing coupled with the novel open access TBpore pipeline to successfully support molecular epidemiology investigations. This was established using well characterized clinical isolates from two and three distinct outbreaks and sublineages (L4, L5, L6) of the *M. tuberculosis* complex with MIRU and Illumina WGS molecular typing results. Further validation of this novel ONT-based clustering approach on larger datasets with increased

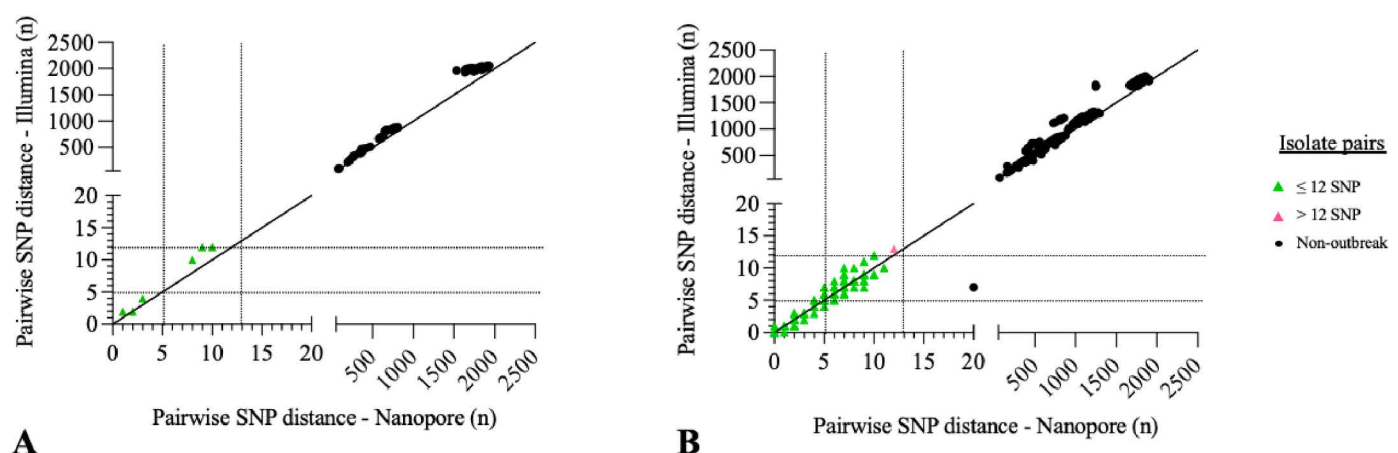


Fig 2. Illumina and Nanopore Whole Genome Sequencing-based clustering comparative analysis. Between isolate pairwise SNP distances using Illumina (COMPASS+psdm; y-axis) and Nanopore (TBpore; x-axis) data for WGS-based typing of *M. africanum* (panel A) and *M. tuberculosis* (panel B) retrospective outbreak isolates. Isolate pairs with identical MIRU-types (triangles) either fall below (green) or above (pink) the 12 SNPs threshold.

<https://doi.org/10.1371/journal.pone.0325914.g002>

geographic catchment, lineage diversity and robust epidemiological metadata should be performed to confirm these results.

Compared to the established standard Illumina NGS platforms, we observed a trend towards lower between isolates SNP distances using ONT and TBpore clusters although, except in one case, this did not affect isolates clustering when using previously accepted transmission thresholds [8,9]. Those between isolates SNP distances differences can be due to the underlying sequencing technology, data pre-processing, processing pipelines, or variant calling parameters [5,10]. Although this finding should also be validated in larger international datasets including samples with more diverse epidemiological and genomic backgrounds, it is consistent with previous work on the validation of Nanopore technology for TB genomic clustering analyses in which this platform was also found to cluster more isolates together [5]. In this previous work, the SNP calling features and clustering pipeline were not the same as those included in TBpore clusters which we evaluated in this study. TBpore uses a more focused decontamination database and optimises variant calling parameters to allow execution of the tool on a standard laptop hence increasing portability of the Nanopore companion analysis pipeline. The previously published analysis used a Nanopore-adapted pipeline which was tailored to a high-performance computing environment where resources were not limiting. Also, clinical epidemiology data from field investigations was not available to contextualize between-sequencing technologies discrepancies.

In our study, the ONT platform and novel pipeline led to the clustering of two isolates which would not have been linked together using standard relatedness thresholds and Illumina technology. Field epidemiology data from public health authorities confirmed that those two isolates were indeed clinically related. Those two isolates also shared the same MIRU-type. Conversely, one pair of isolate was clustered together using Illumina technology but the two isolate did not share an identical MIRU-type and patients had not been identified as being epidemiology related by public health investigations. This potential increased ability of ONT to accurately identify clinically related isolates using previously proposed SNP distance thresholds should be confirmed on larger TB WGS datasets with robust epidemiological metadata.

Methods

Sample selection

Clinical isolates included in this study were retrospectively obtained from the Québec Public Health Laboratory (LSPQ). Isolates were purposively included from two distinct domestic TB outbreaks which had previously been characterized by public health investigations and MIRU-VNTR typing [7]. MIRU-VNTR typing was performed at the Canadian National Microbiology Laboratory (LMN). Case isolates (*M. africanum*; n=4, *M. tuberculosis*; n=21) were epidemiologically related and control isolates (*M. africanum*; n=12, *M. tuberculosis*; n=21) were not. All control isolates were matched with cases based on geographic (administrative region) and temporal (year of sampling) criteria. Reference *M. africanum* (ATCC 25420) and *M.tuberculosis* H37Rv (ATCC 27294) strains were also included as controls for the bioinformatics analyses.

Culture, extraction, DNA sequencing and quality control

Culture and DNA extraction were performed as per standard protocols on MGIT liquid culture media systems (Becton Dickinson, Franklin Lakes, NJ, USA). All isolates were sequenced with both Illumina and ONT platform from the same DNA extractions. Extracts were sequenced on the Illumina NextSeq500 platform at the Institute for Research in Immunology and Cancer (IRIC). A Paired-End library (Kapa Hyperprep Plus) was prepared on-site following the manufacturer's instructions and loaded onto a Mid Output flow cell. ONT sequencing libraries were prepared using the ligation sequencing kit 1D (SQK-LSK109) and the Native Barcoding kits 1D (EXP-NBD104 and EXP-NBD114) following the manufacturer's instructions. The libraries were sequenced on R9.4.1 flow cells on the MinION platform. Detailed sequencing results metrics including genomic coverage are presented as S3 in [S1 File](#).

Molecular typing and phylogenetic analyses

The Illumina FASTQ data had contaminant (non-*M. tuberculosis*) reads removed using Clockwork [11]. Decontaminated FASTQ files were used to identify genomic single nucleotide polymorphisms (SNP) and to generate the consensus sequences with CompassCompact (V1.0.2). Briefly, CompassCompact maps reads to the reference with Stampy and calls SNPs with samtools and bcftools. A pairwise distance matrix was created using psdm (V0.1.0) [12]. Nanopore fast5 data were basecalled, filtered, and demultiplexed using Guppy software (V6.1.5) using the super-accuracy model. The resulting FASTQ files were then processed through TBpore (V0.6.0) subcommand *process* in order to characterize the phylogenetic lineage, call variants and generate consensus sequences [5,13]. This removes contamination from the reads using minimap2 against a reduced version of the database used by Clockwork above. It then maps the reads to the reference using minimap2 and calls variants using bcftools. Exact variant filters were previously described in Hall *et al.* [4]. The consensus sequence was obtained by applying all SNPs that passed filtering to the reference genome and masking those that failed filtering (replace with N). The distance between isolates was determined by the consensus sequences using subcommand *cluster*, which uses psdm to generate a SNP distance matrix from the Nanopore consensus sequences. In both ONT and Illumina pipelines, *M. tuberculosis* H37Rv (NCBI NC000962.3) was used as reference for alignment and SNP calling. Lineages were called for both Illumina and Nanopore data using Mykrobe, which uses the lineage-specific SNP markers as previously described [14].

Clustering comparative analysis

A pairwise distance matrix was used to compare SNP distances between isolates for both NGS technologies. As per previously proposed thresholds, all pairs of isolates having an SNP distance within 5 and 12 were respectively considered as probable or possible transmission clusters [8]. Those clustering results were compared with the MIRU-VNTR genotyping data and contextualized with insights from conventional epidemiological investigation [7]. A paired permutation statistical test was used to assess the statistical significance of the observed differences in SNP distances between technologies. This test was selected as the SNP distance differences between the two technologies represent dependent and paired data points which are not distributed symmetrically around their distribution median.

Statistical analyses

All statistical analyses were performed using R software V 4.3.1.

Ethics

This study was approved by the CRCHUM Institutional Review Board (#2020-8441,19.345). This study was performed on retrospective archived samples. Authors could not identify individual participants.

Supporting information

S1 File. Supporting information 1. MIRU-VNTR and WGS typing results of the *M. tuberculosis* subsp. *tuberculosis* lineage 4 outbreak. Supporting information 2. Statistical Analysis of Illumina Data Compared to Nanopore Data. Supporting information 3. Illumina and Nanopore Sequencing Performance Metrics. (DOCX)

Acknowledgments

We thank the Laboratoire de Santé Publique du Québec (LSPQ) for contributing to the study isolates. We thank the clinical and epidemiology teams who conducted the detailed investigations.

Author contributions

Conceptualization: Sophie Gagnon, Simon Grandjean Lapierre.

Data curation: Sophie Gagnon, Floriane Point, Michael B. Hall, Simon Grandjean Lapierre.

Formal analysis: Sophie Gagnon, Emmanuelle Ametepe, William Cloutier Charette, Pierre-Marie Akochy, Hafid Soualhine.

Funding acquisition: Simon Grandjean Lapierre.

Investigation: Arpita Chakravarti, Pierre-Marie Akochy, Paul Rivest.

Methodology: Sophie Gagnon, Simon Grandjean Lapierre.

Project administration: Simon Grandjean Lapierre.

Resources: Simon Grandjean Lapierre.

Software: Zamin Iqbal, Michael B. Hall.

Supervision: Simon Grandjean Lapierre.

Visualization: Sophie Gagnon, Michael B. Hall, Simon Grandjean Lapierre.

Writing – original draft: Sophie Gagnon.

Writing – review & editing: Emmanuelle Ametepe, Floriane Point, William Cloutier Charette, Arpita Chakravarti, Pierre-Marie Akochy, Paul Rivest, Hafid Soualhine, Zamin Iqbal, Michael B. Hall, Simon Grandjean Lapierre.

References

1. Walker TM, Merker M, Knoblauch AM, Helbling P, Schoch OD, van der Werf MJ, et al. A cluster of multidrug-resistant *Mycobacterium tuberculosis* among patients arriving in Europe from the Horn of Africa: a molecular epidemiological study. *Lancet Infect Dis*. 2018;18(4):431–40. [https://doi.org/10.1016/S1473-3099\(18\)30004-5](https://doi.org/10.1016/S1473-3099(18)30004-5) PMID: 29326013
2. 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;52(10):3795–8. <https://doi.org/10.1128/JCM.01726-14> PMID: 25078914
3. Walker TM, Cruz ALG, Peto TE, Smith EG, Esmail H, Crook DW. Tuberculosis is changing. *Lancet Infect Dis*. 2017;17(4):359–61. [https://doi.org/10.1016/S1473-3099\(17\)30123-8](https://doi.org/10.1016/S1473-3099(17)30123-8) PMID: 28298254
4. Walker TM, Ip CLC, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, et al. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *Lancet Infect Dis*. 2013;13(2):137–46. [https://doi.org/10.1016/S1473-3099\(12\)70277-3](https://doi.org/10.1016/S1473-3099(12)70277-3) PMID: 23158499
5. Hall MB, Rabodoarivelo MS, Koch A, Dippenaar A, George S, Grobbelaar M, et al. Evaluation of Nanopore sequencing for *Mycobacterium tuberculosis* drug susceptibility testing and outbreak investigation: a genomic analysis. *Lancet Microbe*. 2023;4(2):e84–92. [https://doi.org/10.1016/S2666-5247\(22\)00301-9](https://doi.org/10.1016/S2666-5247(22)00301-9) PMID: 36549315
6. Hunt M, Bradley P, Lapierre SG, Heys S, Thomsit M, Hall MB, et al. Antibiotic resistance prediction for *Mycobacterium tuberculosis* from genome sequence data with Mykrobe. *Wellcome Open Res*. 2019;4:191. <https://doi.org/10.12688/wellcomeopenres.15603.1> PMID: 32055708
7. Chakravarti A, Soualhine H, Sharma M, Rivest P, Poirier L, Lavallée C. Indigenous Transmission of *Mycobacterium africanum* in Canada: A Case Series and Cluster Analysis. *Open Forum Infect Dis*. 2019;6(4):ofz088. <https://doi.org/10.1093/ofid/ofz088> PMID: 30949534
8. 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;13:110. <https://doi.org/10.1186/1471-2334-13-110> PMID: 23446317
9. 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. *Lancet Respir Med*. 2014;2(4):285–92.
10. Hall MB, Wick RR, Judd LM, Nguyen AN, Steinig EJ, Xie O, et al. Benchmarking reveals superiority of deep learning variant callers on bacterial nanopore sequence data. *Elife*. 2024;13:RP98300. <https://doi.org/10.7554/eLife.98300> PMID: 39388235
11. Hunt M, Letcher B, Malone KM, Nguyen G, Hall MB, Colquhoun RM, et al. Minos: variant adjudication and joint genotyping of cohorts of bacterial genomes. *Genome Biol*. 2022;23(1):147. <https://doi.org/10.1186/s13059-022-02714-x> PMID: 35791022
12. Hall MB. psdm: Compute a pairwise SNP distance matrix from one or two alignment(s). Available from: <https://github.com/mbhall88/psdm>. 0.1.0 ed: Zenodo; 2021.
13. Hall MB. tbpore. Available from: <https://github.com/mbhall88/tbpore>. 0.7.1 ed: github; 2024.
14. 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. <https://doi.org/10.1371/journal.pone.0041253> PMID: 22911768