Bioinformatics Approaches for Analyzing Short-Read Illumina Sequences of Whole *Mycobacterium tuberculosis* Genomes from an Outbreak

Bhavjinder K. Dhillon[1], Shannan J. Ho Sui[1], Jennifer L. Gardy[2,3], Robert C. Brunham[2], Patrick Tang[2,3], Fiona S.L. Brinkman[1].

[1] Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, British Columbia, Canada
[2] British Columbia Centre for Disease Control, Vancouver, British Columbia, Canada
[3] University of British Columbia, Vancouver, British Columbia, Canada

Corresponding author email: bkdhillon@sfu.ca

An outbreak of tuberculosis occurred over a three-year period in a medium-sized community in British Columbia, Canada. Traditional MIRU-VNTR genotyping suggested a clonal outbreak and traditional contact tracing was unable to identify a source. Whole genome sequencing of 32 *M. tuberculosis* outbreak isolates plus four historical isolates was completed using Illumina-GAII short-read sequencing, in an attempt to describe the outbreak dynamics at a higher resolution. Multiple short-read alignment algorithms (Bowtie, BWA, mrsFAST and SSAHA2) and structural variation prediction methods (BreakDancer, Mauve, VariationHunter and Variation Ascertainment Algorithm) were compared when assessing the sequences. Integrated social network analysis data was used to determine the origins and transmission dynamics of the outbreak. We found that SNP predictions were fairly similar amongst the short-read aligners evaluated, except cases within repetitive sequence regions. Larger insertions/deletions varied more notably as each algorithm proved to have boundaries based on the predictive technique employed. SNP phylogenetic analysis supported two genetically distinct TB lineages suggesting two concomitant outbreaks, and individuals acting as superspreaders. Comparison of short-read alignment tools and structural variation prediction methods revealed capabilities and limitations for each technique. By integrating bacterial whole genome sequencing and social network analysis, more insight into the dynamics of this outbreak was obtained.