DEVELOPMENT OF COMPUTATIONAL TOOLS AND MANAGEMENT OF NEXT GENERATION MYCOBACTERIAL SEQUENCING DATA

Alecia Naidu[1], Peter van Heusden[1], Rob Warren[2], Nico Gey van Pittius[2] & Alan Christoffels[1]

[1]South African National Bioinformatics Institute, University of Western Cape, Modderdam Road, Bellville 7535, South Africa, e-mail: alecia@sanbi.ac.za; *alan@sanbi.ac.za*

[2]DST/NRF Centre of Excellence in Biomedical Tuberculosis Research, US/MRC Centre for Molecular and Cellular Biology, Faculty of Health Sciences - Stellenbosch University

The current advancement in Next generation sequencing technologies, together with the major improvements in computational capabilities, have made it possible to generate and assemble huge microbial datasets. Focusing on *Mycobacterium tuberculosis* (TB), we aim to develop a short-read analysis work-flow for identifying and curating novel polymorphisms in hypo- and hyper-virulent genomes of the T85 strain. This will facilitate the drawing of inferences regarding the relationship between genetic variation and disease in *M. tuberculosis* infection. Current computational tools for handling next-generation sequences have not been exhaustively assessed in microbial genomes. We describe our assessment of tools best-suited to handle TB genome sequences. The work-flow has been implemented using Galaxy which is a user friendly, scalable platform for tool and data integration. Custom made scripts written using the Perl language, for data preprocessing and filtering, have been integrated together with short-read mapping tools to form the analytical steps of the work-flow. Galaxy which is an open-source software is freely available for download. The initial run of the work-flow on the T85 sequence data has generated a total number of 2987 Single Nucleotide Polymorphisms (SNPs). This computational protocol significantly enhances fast analysis and efficient management of huge volumes of microbial sequence data generated using the next generation sequencing technology.