Challenges for detecting genetic variation of *Leishmania donovani* in natural populations and among experimentally induced drug-resistant lines

Hideo Imamura[1], An Mannaert[1], Manu Vanaerschot[1], Tim Downing[2], Matt Berriman[3], Shyam Sundar[4], Suman Rijal[5], Jean-Claude Dujardin[1]

[1]Institute of Tropical Medicine, Antwerpen, Belgium; [2]National University of Ireland, Galway, Ireland; [3]Wellcome Trust Sanger Institute, Hinxton, UK; [4]Baranas Hindu University, Varanasi, India; [5]BP Koirala Institute of Health Sciences, Dharan, Nepal

Next generation sequencing technology (NGS) provides us opportunities to characterize the genome diversity of pathogens in unprecedented details. Whole genome analysis of large numbers of samples, however, still remains challenging partly because of misassemblies and miscalled bases. This occurs even with high quality reference genomes, partly because of uneven depth coverage and SNP bias associated with some sequence runs. In addition, the identification of structural variants (gene or chromosome copy number variation) appears more and more relevant for the understanding of some pathogen's biology. Hence the tasks of correctly identifying genomic variants in NGS reads are challenging. We demonstrate how to handle these challenges using as model *Leishmania donovani*, a parasitic Protozoa (Kinetoplastida). In the frame of a genome diversity project, we created a high quality draft sequence of the parasite and sequenced over 200 clinical isolates and experimentally induced drug-resistant strains. We used a set of sequence analysis tools such as pileup, mpileup, GATK and CORTEX, which uses De Bruijn graph to identify base and structural variations even in the presence of errors in the reference. In addition, we used an allele base distance method to identify critical base change variants that are potentially responsible for drug resistance and also to screen false SNPs among populations. Allele base distance method is particularly instrumental to handle a genome with extensive aneuploidy whose impact on genome variation and on practical SNP calling is still unknown. Our analyses overall indicated that combining many tools and post-analysis data screening are essential to eliminate false positive genomic variants in NGS data sets.