Evaluating the accuracy of genotype imputation in the Major Histocompatibility Complex (MHC) region in selected African populations.

Ruth Nanjala (1), Nicola Mulder (2), Mamana Mbiyavanga (2), Suhaila Hashim (1,3), Santie de Villiers (1,3)

(1) Department of Biochemistry and Biotechnology, Pwani University, Kilifi, Kenya

(2) Computational Biology Division, University of Cape Town, South Africa

(3) Pwani University Biosciences Research Centre (PUBReC), Pwani University, Kilifi, Kenya

Genome wide association studies (GWAS) traditionally use genotyping arrays to genotype large sets of individuals and thus determine which Single Nucleotide Polymorphism (SNPs) are significantly overrepresented in the cases compared to the controls and in this way determine association with disease. Genotyping arrays are cheaper than sequencing but only measure a portion of selected SNPs across the genome. To increase the number of SNPs available, one can use a reference panel of whole genome sequence data from related populations to impute SNPs from those on the array. Some regions in the human genome such as the Major Histocompatibility Complex (MHC) are highly variable and thus difficult to impute. The MHC region in humans has been associated with autoimmune and infectious diseases, adaptive and innate immune responses and adverse responses to organ transplantation. The aim of this study is to therefore evaluate the accuracy of MHC imputation especially in African populations as they have high diversity and an extended linkage disequilibrium. The study sets will be selected from the Gambian individuals within the Gambian Genome Variation Project and simulated data from the South African population. The reference dataset will be chosen from the African populations within the1000 Genome Project, the Gambian sub-population within the 1000 Genome Project and the H3Africa reference panel through their imputation service. Human Leukocyte Antigen (HLA) typing will be done using the OptiType tool. HLA alleles will be imputed from SNP data using HIBAG, SNP2HLA, Minimac4 and IMPUTE2. The assessment metrics will be concordance rate, squared Pearson correlation coefficient and call rate. It is anticipated that the most appropriate software and reference panel for MHC imputation in African populations will be identified. The study will also be able to determine whether sample size and choice of genotyping array influences accuracy and efficiency of MHC imputation.