PP 26 - Implementation of a cloud-based bioinformatics pipeline for genetic variant discovery of next generation sequence data linked to hereditary cancer syndromes.

N. F. Neththikumara[1*], T. K. Wetthasinghe[1], V. H.W. Dissanayake[1]

[1]Human Genetics Unit, Department of Anatomy, Faculty of Medicine, University of Colombo

Introduction: Next generation sequencing (NGS) has now become the norm for medical diagnostics worldwide. Sequencing data used for genetic variant discovery consists of millions of short sequences called reads. Depending on the platform used for sequencing, the reads have different properties like pairing, length and error-rate which can significantly impact the analysis and variant calling process. Thus, analysing NGS data using bioinformatics techniques has become a challenge in terms of complexity, efficiency and cost effectiveness.

Methods: Index cases and at-risk family members of patients with hereditary cancer syndromes have been recruited for studies conducted in the Human Genetics Unit. These individuals have been sequenced by the Illumina Miseq NGS sequencer, with the Trusight Cancer library enrichment kit. An automated bioinformatics pipeline has been developed to analyse these data, which process mapping, filtering, realignment, recalibration, variant calling and annotation. Each module has been integrated together using python and bash scripts. A secondary variant analysis module was also implemented to systematically prioritize variants in genes associated with hereditary cancer syndromes. The pipeline has been deployed to a dedicated cloud server utilizing the Google cloud computing engine and cloud storage services.

Results: This pipeline generates a fully annotated variant calling file (VCF) of single nucleotide polymorphisms (SNPs) and short insertions and deletions (Indels) with several data quality statistics reports. It also incorporates virtual gene panels to generate a variant list in a subset of genes associated with a specific hereditary cancer syndrome, with in-silico pathogenicity predictions for each variant. The amount of sample data that can be run at a given time is presumed to be unlimited, as the pipeline is fully scalable.

Conclusion: The incorporation of custom bioinformatics pipelines into the sequencing workflow enables reliable, efficient and cost-effective data analysis for disease diagnosis.

Keywords: Bioinformatics, Hereditary cancer, Next generation sequencing, Pipeline, VCF

*Email: nilakshafreezon@gmail.com