Evaluation of Effects of Different Base Calling Models on Single Nucleotide Variant Calling Using Low-coverage Long Read Sequencing

Hamza Umut Karakurt hamza_karakurt@windowslive.com

Turkey Idea Technology Solutions R&D Center / Department of Bioengineering, Gebze Technical University, Kocaeli, Turkey

Hasan Ali Pekcan hasan.pekcan@ideateknoloji.com.tr Turkey Idea Technology Solutions R&D Center AyşeKahraman ayse.kahraman@ideateknoloji.com.tr Turkey Idea Technology Solutions R&D Center EsraÇınar esra.cinar@ideateknoloji.com.tr Turkey Idea Technology Solutions R&D Center

Bilçağ Akgün bilcagakgun@gmail.com TurkeyDepartment of Medical Genetics, Faculty of Medicine, Izmir University of Economics

Long-read sequencing technologies such as Oxford Nanopore Technologies (ONT) enabled researchers to sequence long reads fast and cost effectively. ONT sequencing uses nanopores integrated to semiconductor surfaces and sequences the genomic materials using changes in voltage across the surface as each nucleotide passes through nanopore. The default output of ONT sequencers are in FAST5 format. First and one of the most important steps of ONT data analysis is the conversion of FAST5 files to FASTQ files using “basecaller” tools. Generally basecaller tools use pre-trained deep learning models to transform electrical signals to reads. Guppy, the most commonly used basecaller, uses 2 main model types, fast and high accuracy. Since the computation duration is significantly different between these two models, the effect of models on variant calling process has not been fully understood. The aim of this study is to evaluate the effect of different models on performance on single nucleotide variant calling. Here, we used 8 low-coverage long read sequencing results of NA12878 (gold standard data) to compare the variant calling results of Guppy. Each data is basecalled using Guppy fast and high accuracy models and output FASTQ files aligned with minimap2 using human genome (hg19). Variants are called used Clair3 and final VCF files compared using R programming environment. Genome In A Bottle (GIAB) NA12878 high confidence variants file used for true positive variants. Obtained results indicated that pass/fail ratios of base called datasets and computation times are significantly higher in high accuracy models. Also, the number of called variants is remarkably higher in fast models but the true positive variant ratio difference is significantly smaller. The primary observation in our case using fast models does not decrease the ratio of true positive rate but decrease the number of called variants.