AcuRe: Third-Generation Machine Learning Cancer Detection Model Using Nanopore Read Alignments

Pahlavan, Natalia (School: Jericho High School)

Cancer is the cause of every sixth death globally with inadequate screening attributing to 40,000 deaths in the U.S. annually. An alternative to current genetic mutation screening efforts is Oxford Nanopore Technologies' (ONT's) portable MinION device, which conducts low-cost, real-time sequencing of nucleic acids. ONT's NA12878 human reference was mapped against a Zymo microbial standard using Mappy library in C programming language; Minimap2 produced a concerning error rate of ~3%, which, when compared to the human genome, equates to 96,000,000 incorrectly called base pairs. This research aims to 1) discern ONT's variant calling pipeline 2) elucidate Phred quality scores between aligned and misaligned reads and 3) develop a machine learning algorithm, AcuRe, to classify numerical scores associated with read alignment accuracy. Quality scores were extracted from FASTQ files using Python and the first 100 noisy base pairs were trimmed for preprocessing; the quality scores demonstrated 35% of misaligned reads obtained a score of 5, whereas 27% of aligned reads obtained a score of 10. Quality scores were used to conduct a K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) classifier as AcuRe produces an accuracy of 95% and 88% for the SVM and KNN, respectively. Therefore, Minimap2's erroneously aligned reads were reduced by 20x, which increases the accuracy for downstream tools as only correctly aligned reads from AcuRe are processed for follow-up softwares. AcuRe is a free mutation detection software that can be applied to genetic disorders, reducing the $13 billion spent in the U.S. annually.