Best Practices for Bioinformatics Pipelines for Molecular-Barcoded Targeted Sequencing

Marcel Caraciolo[1], Wilder Barbosa Galvao[1], George de Vasconcelos Carvalho Neto[1], Rodrigo Bertollo[1], Joao Bosco Oliveira[1],

1 Genomika

Abstract

In cancer research the detection of mutations is critical, for tumor samples and blood samples mutations may be present in very low fractions of DNA molecules. By using molecular barcoding technology, more than reduce the impact of enrichment, the sequencing errors can be eliminated by tagging each input molecule with an unique molecular identifier (UMI). In contrast to sample barcoding, molecular barcoding assigns a unique sequence not just to all the molecules from a certain sample, but to all molecules being amplified and sequenced. Despite the difference it is common to have both sample barcodes and molecular barcodes in the same sequencing reads. Recent works on this approach show outstanding performance in targeted high-throughput sequencing, being the most promising approach for the accurate identification of rare variants in complex DNA samples, and has application in several areas such as detecting DNA mutations at very low allele fractions with high accuracy for cancer samples and reducing sequencing artifacts occurrences. However, at the sample preparation, the residual PCR errors might be introduced at first PCR cycles and during UMI tag attachment, which decrease the accuracy of variant calling. In order to perform the variant detection on those input data, a different approach is required for bioinformatics pipelines that handles the caveats of UMI-based analysis. By using specific algorithms and softwares, the pipeline is designed to obtain high-fidelity mutation profiles and call ultra-rare variants. In this poster we present the best practices and strategies for handling the UMI-tagged data, by showing the steps and related software tools to the audience when building the variant calling pipeline

Funding: Genomika