The importance of an adequate soft-clip based approach on bioinformatics pipelines for multiplex targeted next-generation sequencing

George Carvalho, Renata Andrade, Marcel Caraciolo, João Bosco Oliveira, Rodrigo Bertollo

Genomika Diagnósticos

Advances in high-throughput sequencing have enabled the adoption of sequencing for various applications in research and clinical diagnostics. In addition to lower per-base sequencing costs, one of the crucial factors in reducing per sample sequencing costs is the ability to focus sequencing throughput on specific target regions of interest. One of the main strategies for accomplishing this goal is the use of PCR-based enrichment method by using a few high dimension multiplex PCR reactions (Ampliseq from Life, GeneRead from QIAGEN). The products of PCR enrichment include the primers on both ends. However, these primers are not native to the sample, and need to be removed before variant calling as not to disturb the variant calls from other amplicons that overlap these primers. There are several methods for primer removal, but depending on the strategy selected it might lead to low coverage of reads at the targeted region or missing variants that are located at near the edge of the reads. In some cases, removing the primers at the raw data (FASTQs) can cause misalignments which can lead to a false-positive calls. In this poster, we show that using soft-clipping of the read bases of the primers instead of removing it, it can improve the variant calling sensitivity. We built a custom pipeline for variant calling for amplicon reads using open-source tools such as Cutadapt, BWA, Picard and GATK. For primers base masking we used the tool KATANA, and we compared our results with another pipeline produced by the primers provider. Preliminary tests, conducted with 73 patients, identified 955 variants that compared to the provider's results yielded 85.23% true-positive, 5.13% false-positive and 9.63% false-negative rates. Among the false-positives, approximately 1% of the variants were true and all false-negatives were the results of bad trimming on the provider's part. We outperformed the results of the provider pipeline reducing the number of false-positives and false-negatives due to incorrect primer masking and missing low coverage variant calls. In this poster we would like to share with the audience the lessons learned during the development and present the best practices and strategies to work with amplicon reads in variant calling pipelines.