[BIOINFO 2025]

Variety-Specific Rice Samples and Their Influence on False Positives in Variant Calling

Sunhee Kim, Dongju Lee, and Chang-Yong Lee

Department of Industrial Engineering, Kongju National University
**Corresponding author: clee@kongju.ac.kr*

In this study, we investigated the effect of variety on false positives in rice genetic variant calling from next-generation sequencing data. To this end, we used two major rice genome strains, Indica and Japonica, and their reference sequences to develop variant calling models with different sample compositions from the two varieties. We found that the greater the difference between the samples and the reference sequence, the more variants were predicted. We constructed confusion matrices from the predicted variants to understand this finding and compare the performance of different variant calling models, using them as performance measures in machine learning. In particular, We found that a significant proportion of the incrementally predicted variants are potential false positives, a trend that becomes more pronounced the more phylogenetically different accessions from the reference are included in the samples. To improve the accuracy of the predicted variants, we proposed a method to identify false positives that can be excluded if necessary. We demonstrated the validity of this method using different variant calling models, showing a reduction in false positives in the predicted variants.