A Method of Identifying False Positives in the Variant Calling

Sunhee Kim, Dongju Lee, and Chang-Yong Lee
Department of Industrial and Systems Engineering, Kongju National University

In this study, we investigated the strain-specific effect in genetic variant calling from next-generation sequencing data. For this purpose, we used two major strains of the rice genome, Indica and Japonica, to build different variant calling models that differ in the composition of samples from the two strains. We found that the more the samples differed in their strains from the reference sequence, the more variants were predicted. We used machine learning approaches to understand this finding and compared the performance of different variant calling models using confusion matrices constructed from the predicted variants. We found that a significant proportion of the incrementally predicted variants are potential false positives, which becomes more pronounced the more phylogenetically different accessions from the reference are included in the samples. For the accuracy of the predicted variants, we proposed a method to identify the false positives that can be excluded from the potential false positives if necessary. The proposed method involves calling true variants from the purebred samples. We demonstrated the validity of the proposed method on the different variant calling models and showed a reduction of false positives in the predicted variants

Keywords: Variant calling, False positives, Reference sequence, Machine learning