The role of known genetic variants in the base quality score recalibration for the genetic variant calling

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

The genetic variant calling from genome data relies on various pipelines of computational tools to account for systematic differences in the genome data of different species. In particular, the base quality score recalibration (BQSR) in the pipeline is a pre-processing step that leverages a large database of known variants called dbSNP. While these pipelines are expected to be applicable in a species-independent manner, they have not been carefully evaluated with non-human data. To investigate the impact of the dbSNP on BQSR, we analyzed genomic sequencing data from four different species: human, sheep, rice, and chickpea. Unexpectedly, the recalibrated scores and the error rate obtained by BQSR were biased by the size of the dbSNP and its builds. To address this issue, we suggest an alternative to the dbSNP by constructing a pseudo-database for various species based on the sequence data.

Keywords: Base quality score recalibration, dbSNP, Genetic variant calling, Error rate, Genome data