[BIOINFO 2024]

GVRP: Genome Variant Refinement Pipeline for Variant Analysis in Non-Human Primates Using Machine Learning

Jeong-Hoon Choi[1], Kibeom Kim[1], Jonghyeon Bae[1], and Giltae Song[1,2,3*]

[1]*Division of Artificial Intelligence, Department of Information Convergence Engineering, Pusan National University*

[2]*Center for Artificial Intelligence Research, Pusan National University*

[3]*School of Computer Science and Engineering, Pusan National University*

**Corresponding author: gsong@pusan.ac.kr*

Numerous inquiries into human diseases rely on model systems, including non-human primate, alongside their comprehensive genome analyses. While DeepVariant, a genome analysis pipeline that uses a deep neural network, excels in calling human genetic variations, its reliance on calibrating against a set of known variant sites from previous population studies poses challenges for non-human primate. In this work, we introduce the Gevnome Variant Refinement Pipeline (GVRP), employing a machine learning-based approach to improve the accuracy of variant detection in non-human primate to address above limitation. Rather than training separate variant callers for each species, we employ a machine learning model to accurately identify variations and filter out false positives from DeepVariant. In GVRP, we omit certain DeepVariant preprocessing steps that do not apply for non-human primate genome data and leverage the ground-truth Genome In A Bottle (GIAB) variant calls to train the machine learning model for non-human primate genome variant refinement. The evaluation of our refinement pipeline demonstrates outstanding performance in both quality and quantity aspects for both human and non-human primate genome variants. GVRP is the machine learning based variant refining model that presents a robust toolkit for enhancing the accuracy of variant calls in non-human primate. We anticipate that GVRP will significantly expedite genome variation studies for non-human primate.