AIVariant1: a deep learning-based somatic variant detector for highly contaminated tumor samples

Hyeonseong Jeon[1,2,*], Junhak Ahn[2,3,*], Byunggook Na[4], Soona Hong[5], Lee Sael[6], Sun Kim[7], Sungroh Yoon[4,8], and Daehyun Baek[1,2,3,8,#]

[1]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 08826, Republic of Korea.

[2]Genome4me Inc., Seoul 08826, Republic of Korea.

[3]School of Biological Sciences, Seoul National University, Seoul 08826, Republic of Korea.

[4]Department of Electrical and Computer Engineering, Seoul National University, Seoul 08826, Republic of Korea.

[5]AIGENDRUG Co., Ltd., Seoul 08826, Republic of Korea.

[6]Department of Software and Computer Engineering, Ajou University, 16499, Suwon, Republic of Korea.

[7]Department of Computer Science and Engineering, Seoul National University, Seoul 08826, Republic of Korea.

[8]Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul 08826, Republic of Korea.

[#]Correspondence should be addressed to D.B. (baek@snu.ac.kr).

[*]These authors contributed equally to this work.

Abstract

The detection of somatic DNA variants in tumor samples with low tumor purity or sequencing depth remains a daunting challenge despite numerous attempts to address this problem. In this study, we constructed a substantially extended set of actual positive variants originating from a wide range of tumor purities and sequencing depths, as well as actual negative variants derived from sequencer-specific sequencing errors. A deep learning model named AIVariant1, trained on this extended dataset, outperforms previously reported methods when tested under various tumor purities and sequencing depths, especially for low tumor purity and sequencing depth.