Accurate and fast detection of copy number variations from short-read whole-genome sequencing with deep convolutional neural network

Jiajin Li[1,▽], Stephen Hwang[2,▽], Luke Zhang[3], Jae Hoon Sul[4,*]

[1]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles
[2]Department of Biomolecular Engineering, Jack Baskin School of Engineering, University of California, Santa Cruz
[3]Interdepartmental Program for Neuroscience, University of California, Los Angeles
[4]Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles
[▽]These authors contributed equally
[*]Correspondence: jaehoonsul@mednet.ucla.edu

ABSTRACT: Copy number variation (CNV) is a type of genetic mutation where a stretch of DNA is lost or duplicated once or multiple times. CNVs play important roles in the development of diseases or complex traits. It is challenging to detect CNV with short-read DNA sequencing technology because of its significantly varying size. Many methods have been developed but still yield unsatisfactory results with high computational costs. Here, we propose CNV-Net, a novel approach for CNV detection using a six-layer convolutional neural network. We encode DNA sequencing information into RGB images and train the convolutional neural network with these images. The fitted convolutional neural network can then be used to predict CNVs from DNA sequencing data. We benchmark CNV-Net with four high-quality whole-genome sequencing datasets available in Genome in a Bottle Consortium, which are considered as gold standard benchmarking datasets for CNV detection. We demonstrate that CNV-Net is more accurate and efficient in CNV detection than current tools.