Clairvoyante:A Multi-task Convolutional Deep Neural Network for Variant Calling in Single Molecule Sequencing

Ruibang LUO

Department of Computer Science, The University of Hong Kong, Hong Kong

Email: rbluo@cs.hku.hk

The accurate identification of DNA sequence variants is an important but challenging task in genomics. It is particularly difficult for single molecule sequencing, which has a per-nucleotide error rate of ~5%-15%. Meeting this demand, we developed Clairvoyante, a multi-task five-layer convolutional neural network model for predicting variant type (SNP or indel), zygosity, alternative allele and indel length from aligned reads. For the well-characterized NA12878 human sample, Clairvoyante achieved 99.73%, 97.68% and 95.36% precision on known variants, and 98.65%, 92.57%, 77.89% F1-score for whole-genome analysis, using Illumina, PacBio, and Oxford Nanopore data, respectively. Training on a second human sample shows Clairvoyante is sample agnostic and finds variants in less than two hours on a standard server. Furthermore, we identified 3,135 variants that are not yet indexed but are strongly supported by both PacBio and Oxford Nanopore data. Clairvoyante is available open-source (https://github.com/aquaskyline/Clairvoyante), with modules to train, utilize and visualize the model.