DOI 10.18699/MM-HPC-BBB-2018-25

Assessment of software for somatic single nucleotide variant identification using simulated whole-genome sequencing data of cancer

W. Kittichotirat[1]*, P. Khongthon[1], K. Kusonmano[2], S. Cheevadhanarak[2]
[1] *Pilot Plant Development and Training Institute, King Mongkut's University of Technology Thonburi, Bangkok, Thailand*
[2] *School of Bioresources and Technology, King Mongkut's University of Technology Thonburi, Bangkok, Thailand*
** e-mail: weerayuth.kit@kmutt.ac.th*

Key words: analytical pipeline, somatic single nucleotide variants, whole-exome sequencing

Motivation and Aim: Next-generation sequencing is an important tool for identifying disease-causing mutations in human. However, it can be relatively difficult to identify many true somatic single nucleotide variants (sSNVs) because they may not be supported by enough sequencing reads to pass the minimal criteria [1]. This could be caused by the contamination of normal cells, cell population heterogeneity, or sample preservation. As a result, some sSNVs calling software that performs well in one sample may perform poorly in another [2].
Methods and Algorithms: The reference human genome sequence and known variations were used to build a model for generating germline and somatic mutations. This was then used to simulate 2x101 bp paired-end whole-genome sequencing reads at 50x coverage for both normal and cancer samples. The contaminated cancer samples were also constructed at different levels of purity. Four software (VarScan2, SomaticSniper, Strelka, and MuTect) were used to identify sSNVs. The accuracy of variant calling software was assessed by comparing to the known variants using sensitivity and specificity analysis. Moreover, the assessment was also conducted for different combinations of variant calling software to search for the most optimal combination for identifying sSNVs.
Results: VarScan2 had the lowest accuracy in identifying sSNVs in the low purity sample. However, VarScan2 also outperforms all other software in identifying sSNVs in high purity sample. On the other hand, MuTect excelled at identifying sSNVs in low purity sample with sensitivity greater than VarScan2 by 20 folds. Interestingly, the combination of MuTect, Strelka, and VarScan2 provided the highest sensitivity for identifying sSNVs in both pure and contaminated cancer samples.
Conclusion: This study provides an assessment of software for sSNVs identification using simulated cancer and matched normal datasets. The results suggested that combination of outputs from multiple software can help to improve the prediction accuracy.
Acknowledgements: Supported by the Bioinformatics and Systems Biology Program, KMUTT and BIOTEC, NSTDA, Thailand.

References

1. Fang L.T. et al. (2015) An ensemble approach to accurately detect somatic mutations using SomaticSeq. Genome Biology. 16(1):197-209.
2. Ewing A.D. et al. (2015) Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. Nature Methods. 12(7):623-630.