Analysis of sequence-based copy number variation detection tools for cancer studies

Analysis of sequence-based copy number variation detection tools for cancer studies

Sheida Nabavi, Zhengqiu Cai and Peter J Tonellato
Center for Biomedical Informatics, Harvard Medical School

Background: Copy number variations (CNVs) can play an important role in tumor genesis and growth through amplification of oncogenes and loss of tumor suppressor genes. Consequently, identification of cancer specific somatic CNVs can provide insight into cancer diagnosis, prognosis and treatment. Recent advances in high throughput next generation sequencing (NGS) technologies have created an opportunity for detecting CNVs with higher accuracy and resolution than previous technologies such as array-comparative genomic hybridization. As a result, several sequence-based CNV detection methods and software applications have been developed to take advantage of these technologies. However, no comprehensive comparison of the sensitivity and precision of the most recent sequence-based tools has been performed. The results of this work indicate the performance characteristics of the recent sequence-based CNV detection tools, which can facilitate selection of an appropriate tool for cancer studies and serve as a guide to develop new algorithms that address current limitations.

Results: In this work, we analyze the performance of six new publically available sequence-based CNV detection tools (SegSeq, CNVer, CNVnator, FREEC, CNV-seq and ReadDepth), using six synthesized datasets, characterized by different coverage values (0.5, 2, 5, 10, 20 and 50) and contained known CNVs, and eight low coverage breast cancer cell line (MCF7, T47D, BT474, ZR75-1, BT20, MDA-MB-231, MDA-MB-468 and HCC1143) NGS datasets. The synthesized NGS datasets are used to accurately compute sensitivity (Figure 1a), precision and breakpoint accuracy of the tools while increasing sequence coverage. Breast cancer cell line NGS datasets are used to calculate sensitivity and precision of the tools (Figure 1b), analyze statistical characteristics of the detected CNVs and compare the computational requirements and costs of the tools employing default and a range of the key parameters' values.

Conclusion: The sensitivity and breakpoint accuracy of CNV detection tools, using synthesized and breast cancer cell lines NGS datasets, indicate tools that employ more computationally complex and elaborate detection algorithms; and tools that incorporate information embedded in paired-end data and depth of coverage, perform better than other tools (in average 30% increase in sensitivity and 6% increase in break point accuracy. However, the computational cost of the better performing tools is higher than the other methods (about one order of magnitude). Also, this study shows that the precision of current CNV detection tools is still low (43% in average) and indicates the likelihood of high false positive rate. There is also a disturbing lack of CNV prediction consensus across tools – even those high performing methods. Current methods for detecting CNV by analysis of NGS sequence of cancer genomes do not yet provide the level of accuracy and robustness required to predict CNVs with high confidence. This study demonstrates the need for further work on both the quality of NGS data and on the development and refinement of sequence-based CNV detection algorithms.

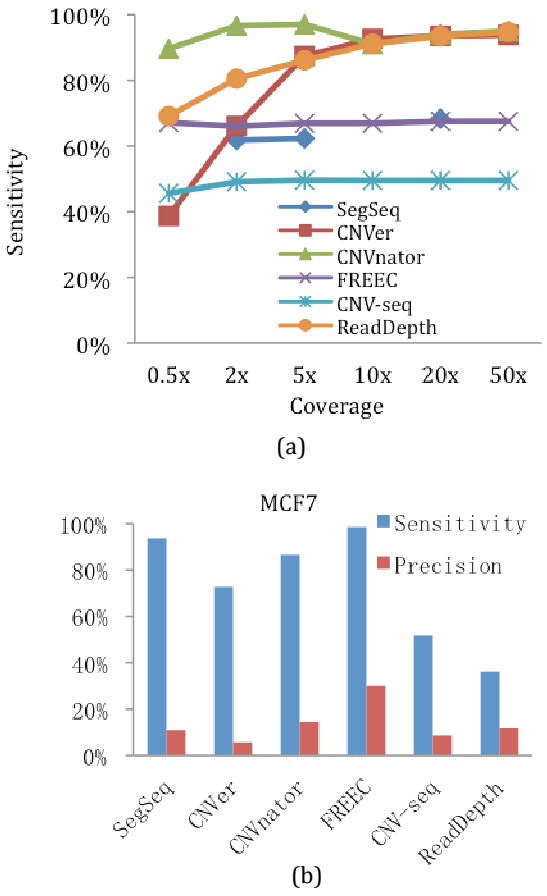

Figure 1. (a) Sensitivity of the tools vs. coverage using synthesized NGS datasets. (b) Sensitivity and precision of the tools against MCF7 cell line