Computational Method for Tumor Cell Detection in Single-Cell DNA Sequencing Data

Toluwanimi Ariyo, Department of Research, Biomedical Sciences Magnet, Ridge View High School

Department of Computer Science, University of Illinois-Urbana Champaign

National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health

Key Words: Single-Cell DNA sequencing, Tumor, Cancer, Machine Learning

ABSTRACT

Single-cell DNA sequencing (scDNA-seq) helps researchers study the evolutionary process of cancer. It is a process used to examine individual cells, describe intra-tumor heterogeneity, and reconstruct the evolutionary history of a tumor. Coverage is the number of reads at a given position in the genome. The depth of high-coverage scDNA-seq allows for analysis of point mutations while it is difficult to make these inferences within low-coverage scDNA-seq. However, due to the uniformity of coverage, ultra-low coverage scDNA-seq is ideal for copy number calling [6].

This study aims to develop a computational method, utilizing features computed from low-coverage scDNA-seq, to detect tumor cells and assist in future efforts of identifying technical errors. Data was pre-processed using Principal Component Analysis (PCA). A machine learning algorithm was implemented to detect tumor cells in this latent, dimensionally reduced space for two patients (patients S0 and S1) with breast cancer sequenced using 10x genomics. The training set (patient S0) had an accuracy of 98% for tumor cell detection. The testing set (patient S1) had an accuracy of 99% for tumor cell detection. This demonstrates that these features are useful for accurately detecting tumor cells in ultra-low coverage scDNA-seq data.