Title

Predictive modeling of cancer stage and location in colorectal cancer patients using transcriptomic data

Authors

Robert Diaz,[1] Brian Grajeda,[2] and Sourav Roy[1,2]

[1]Bioinformatics Program and [2]Department of Biological Sciences, The University of Texas at El Paso, El Paso, TX

Abstract

Colorectal cancer (CRC) was ranked as the third most prevalent cancer globally in 2020, with nearly 1.93 million new cases. In the United States, ethnicity significantly influences the incidence and mortality rates of CRC; notably, Hispanic populations exhibit lower early detection rates compared to Non-Hispanic Whites (NHW). In 2018, CRC accounted for 11% and 9% of all cancer-related fatalities in Hispanic men and women, respectively, alongside 12% and 8% of all new cancer diagnoses. Early detection of CRC elevates survival rates by 90%; however, merely 39% of CRC patients are diagnosed early, largely due to inadequate screening, especially among Hispanics. Limited healthcare access further exacerbates this disparity. Consequently, Hispanics often face advanced-stage CRC diagnoses, correlating with diminished survival rates. In this study, transcriptomic data from 13 Hispanic and 15 NHW CRC patients were generated using high-throughput sequencing platforms like Illumina. Raw data were processed to attain sequenced reads via CASAVA base recognition, stored in FASTQ format. The reads were then aligned and annotated using the Rsubread library against a human genomic reference from NCBI. The generated transcriptomic counts, alongside variable descriptors for each sample, formed the basis of our analytical dataset. Utilizing the tidymodels library in R, the clean data were partitioned into testing (25%) and training sets (75%), followed by cross-validation. Through specifying the target variable and predictors, a Random Forest model facilitated a 95% prediction accuracy in determining cancer stage and location, showcasing a promising avenue for enhancing early CRC detection and intervention, particularly among disadvantaged ethnic groups.