Optimizing Sample Size for Statistical Learning with Bulk Transcriptomic Sequencing: A Learning Curve Approach

Yunhui Qi[1,2], Xinyi Wang[1,3], Li-Xuan Qin[1,*]

[1]Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, United States
[2]Department of Statistics, Iowa State University, Ames, IA, United States
[3]Department of Statistics, The University of California, Davis, CA, United States
*Corresponding Author

Abstract

Accurate sample classification using transcriptomics data is crucial for advancing personalized medicine. Achieving this goal necessitates determining a suitable sample size that ensures adequate statistical power without undue resource allocation. Current sample size calculation methods rely on assumptions and algorithms that may not align with modern machine and deep learning techniques for sample classification. Addressing this critical methodological gap, we present a novel computational approach that establishes the power-versus-sample-size relationship by employing a data augmentation strategy followed by fitting a learning curve. We comprehensively evaluated its performance for microRNA and RNA sequencing data, considering diverse data characteristics and algorithm configurations, based on a spectrum of evaluation metrics. To foster accessibility and reproducibility, the Python and R code for implementing our approach is available on GitHub. Its deployment will significantly facilitate the adoption of statistical learning in transcriptomics studies and accelerate their translation into clinically useful classifiers for personalized treatment.