Removing Unwanted Variation from High-throughput Single-Cell RNA-seq Data

Due to advances in technology, high-throughput RNA measurement at single-cell level is now available. More and more studies are using this technology, which has already led to identification of novel cell types and guiding immunotherapy for cancer treatment. However, this data is highly-complex with a considerable amount of unwanted technical variation that can obscure our ability to detect important biological signals. Current methods for removing these unwanted variations exist but they are not always effective as they tend to remove biological signals whilst trying to remove the unwanted variation. We develop a statistical method for removing unwanted variation from single-cell RNA sequencing data. The statistical method models the raw sequencing count using zero-inflated negative binomial (ZINB) distribution. Using technical replicates and control features (genes), the method estimates the unwanted factors via a modified iterative reweighted least squares (IRLS) algorithm and remove the effect of these unwanted factors from the raw sequencing count using randomized quantile residuals approach. Using simulated and real datasets, we compare the method to leading methods for removing unwanted variation in single-cell RNA sequencing data and demonstrate the comparative advantage of our method in removing unwanted factors while retaining important biological signals. The method is implemented as an R package and available from the following Github site: https://github.com/limfuxing/ruvIIInb