Normalization of generalized transcript degradation improves accuracy in RNA-seq analysis

RNA-sequencing (RNA-seq) is a powerful high-throughput tool to profile transcriptional activities in cells. The observed read counts can be biased by various factors such that they do not accurately represent the relative abundance of the mRNA transcripts. Normalization is an essential step to correct for such biases to ensure fair comparisons of gene expression between samples or conditions. Here we show that the gene-specific heterogeneity of transcript degradation pattern presents a common and major source that may substantially bias the results in differential expression analysis. Most existing normalization approaches focused on global adjustment of systematic biases are ineffective to correct for this bias. We propose a novel method based on nonnegative matrix factorization over-approximation that allows quantification of transcript degradation of each gene within each sample. The estimated degradation index scores are used to build a pipeline named DegNorm (stands for degradation normalization) to adjust read counts for transcript degradation heterogeneity on a gene-by-gene basis while simultaneously controlling the sequencing depth. The robust and effective performance of this method is demonstrated in an extensive set of real RNA-seq data and simulated data.