Removing technical variability in RNA-seq data using conditional quantile normalization

KASPER D. HANSEN, RAFAEL A. IRIZARRY, ZHIJIN WU*

The ability to measure gene expression on a genome-wide scale is one of the most promising accomplishments in molecular biology. Microarrays, the technology that first permitted this, were riddled with problems due to unwanted sources of variability. Many of these problems are now mitigated, after a decade worth of statistical methodology development. The recently developed RNA sequencing (RNA-seq) technology has generated much excitement in part due to claims of reduced variability in comparison to microarrays. However, we find RNA-seq data to demonstrate unwanted and obscuring variability similar to what was first observed in microarrays, in some cases worse. In particular, we find GC-content has a strong sample specific effect on gene expression measurements that, if left uncorrected, leads to false positives in downstream results. We also report on commonly observed data distortions that demonstrate the need for data normalization. Here we describe statistical methodology that improves precision by 42% without loss of accuracy. Our resulting conditional quantile normalization (CQN) algorithm combines robust generalized regression to remove systematic bias introduced by deterministic features such as GC-content, and quantile normalization to correct for global distortions.