PLNseq: A multivariate Poisson lognormal distribution for high-throughput correlated RNA-seq read counts

Hong Zhang*
Fudan University, Shanghai, P. R. China zhanghfd@fudan.edu.cn

Jinfeng Xu
National University of Singapore, Singapore

Xiaohua Hu
Fudan University, Shanghai, P. R. China

Zewei Luo
Fudan University, Shanghai, P. R. China

High-throughput RNA sequence technology provides an attractive platform for gene expression analysis. In many experimental settings, read counts are measured from matched samples or taken from the same subject under multiple treatment conditions. The inherent correlation therefore should be evaluated and taken into account in deriving tests of differential expression. The existing methods either ignore or indirectly model the correlation. Consequently, application of them often results in reduced power or increased false discoveries. We use a multivariate Poisson lognormal distribution (PLNseq) to model correlated read count data. The correlation is directly modeled through the random effects and evaluated by the likelihood principle. A fast three-stage algorithm is further developed for its numerical implementation. Results using simulated data demonstrate that our method performs very well in terms of both power and robustness. Application to two real datasets also returns more meaningful p-values and uncovers more differentially expressed genes than the existing approaches.

Key Words: Differential expression, RNA-seq, Poisson lognormal distribution, correlated sample