“Using Quasi-Likelihood Analysis of RNA-Seq Data to Identify Differentially Expressed Genes”

Dan Nettleton
Laurence H. Baker Endowed Chair
in Biological Statistics and Professor
Department of Statistics
Iowa State University

High-throughput DNA sequencing technologies can be used to identify sequences of bases that occur in samples of RNA. This approach (known as RNA sequencing or RNA-seq for short) provides counts that serve as measures of transcript abundance in a biological sample for each of thousands of genes. The amount of messenger RNA (mRNA) produced by a gene is often referred to as a gene's expression level. Thus, RNA-seq provides expression level measurements for thousands of genes. When RNA-seq technology is applied to multiple independent samples of different types, researchers often want to determine which genes are differentially expressed, i.e., which genes have mean levels of expression that differ across sample types. This talk will describe quasi-likelihood inference strategies for identifying differentially expressed genes. A key step in quasi-likelihood analysis is the estimation of a dispersion parameter which plays a role analogous to that of error variance in linear model analysis. Gene-specific dispersion parameter estimation can be challenging when few observations are available for each gene. To address this challenge, we propose a hierarchical model for dispersion parameters that allows for improved estimation by borrowing information across genes. The approach is implemented in the R package QuasiSeq and described in a recent paper by Lund, Nettleton, McCarthy, and Smyth.