High-Thoughput RNA Sequence Data as a Composition

Dominic LaRoche, Bonnie Lafluer, and Dean Billhiemer

Monday, March 28, 2016

High-throughput, RNA sequence data generated from next-generation sequencers such as those from Illumina and Roche is generally modelled as count data arising from a Poisson or Negative-binomial distribution (many cites). However, these instruments have a maximum number of reads available per run. For example, the Roche 454 GS Junior (tm) claims approximately 100,000 reads per run for shotgun sequencing and 70,000 reads per run for amplicon sequencing. These reads are distributed accross all of the samples included in the run and leads to a total sum constraint on the data. This constraint cascades down to each sample which is, in turn, constrained by the total number of reads allocated to each sample.