Benchmarking ENCODE RNAseq data

Angelika Merkel
Centre de Regulació Genòmica (CRG) , Barcelona, Spain
angelika.merkel@crg.eu

Within the ENCODE project we have generated a matrix to evaluate the consistency of RNAseq experiments based on the correlation of biological replicates with each other. In addition to classic read statistics that can be used to assess library complexity and identify potential amplification artifacts, as well as mapping statistics, we employ a quantitative reproducibility score, the so called "Irreproducible Discovery Rate (IDR)". The statistical method is based on a curve fitting approach with a copula mixture model that partitions the data into a noise and a reproducible group from which the IDR is derived, analogous to FDR.

Our matrix is intended as a users guide to the ENCODE RNAseq data and will be published together with the raw data through the UCSC data portal. It has already aided in the identification of datasets of low consistency and helped downstream analysis.