CountClust: R package for clustering and visualization of RNA-seq data

Kushal K Dey, Matthew Stephens

January 31, 2016

keywords

RNA-seq, clustering, topic model, visualization, batch effects

Abstract

RNA sequencing of both bulk and, more recently, single cells, have become the method of choice for measuring gene expression. The data from these assays are often summarized by counts of the number of reads mapping to different genes. One common analysis step is to cluster the samples, usually using a hierarchical clustering method. Here we explore an alternative: the use of a model-based clustering method, "latent Dirichlet Allocation", previously developed for Natural Language Processing (Blei, Ng and Jordan 2003), that takes account of the count nature of the data. This model, like the the admixture model in population genetics (Pritchard, Stephens and Donnelly 2000), allows that each sample may belong to more than one cluster. We suggest different ways to visualize results, and implement methods to help interpret the clusters by identifying genes whose expression characterizes each cluster. We illustrate the performance and potential of the method by applying it to both the Genotype Tissue Expression (GTEx) Project bulk-RNA data, and to single cell RNA-seq datasets. We also discuss the important issue of batch effects, and how they can influence results. Building on the maptpx package (Taddy 2014) for model fitting, our methods are implemented in an R package, CountClust.