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

With the advent of next generation sequencing technologies, bulk RNA-seq and more recently, single cell RNA-seq (scRNA-seq) experiments have become popular sequencing protocols. The data obtained from these experiments are counts data, representing the number of reads mapping to different genes. We propose a clustering model aimed specifically at modeling counts data. The model is similar to the topic model in Natural Language Processing (Blei, Ng and Jordan 2003) and admixture model in population genetics (Pritchard, Stephens and Donnelly 2000). For model fitting, we use the *maptpx* R package due to Matt Taddy (Taddy 2014). We also suggest visualization tools for interpretable representation of the clusters. We also address the issue of gene set selection for the clustering framework and how to separate the variation due to technical effects or batch effects from the biological variation of interest. We discuss applications of our method to the Genotype Tissue Expression (GTEx) Project bulk-RNA data taken across multiple tissue samples, as well as single cell datasets. The methods and materials discussed have been implemented in our **R** package, *CountClust*.