

S19 Fig. Comparison between GoM model and hierarchical clustering under different scenarios of data transformation. We used GTEx V6 data for model performance comparisons. Specifically, for every pair of the 53 tissues, we assessed the ability of the methods to separate samples according to their tissue of origin. The subplots of heatmaps show the results of evaluation under different scenarios. Filled squares in the heatmap indicate successful separation of the samples in corresponding tissue pair comparison. (a) Hierarchical clustering on log2 counts per million (CPM) transformed data using Euclidean distance. (b) Hierarchical clustering on the standardized log2-CPM transformed data (transformed values for each gene was mean and scale transformed) using the Euclidean distance. (c) GoM model of $K = 2$ applied to counts. (d) Hierarchical clustering on counts data with the assumption that, for each gene the sample read count c_{ng} has a variance $\bar{c}_g + 1$ that is constant across samples. And, the gene-specific variance $\bar{c}_g + 1$ was used to scale the distance matrix for clustering. (e) Hierarchical clustering applied to adjusted count data from c(c). Each gene expression is further normalized over (c) to have mean expression value of 0 and variance of 1. Taken together, these results suggest that regardless of the different data transformation scenarios, the GoM model with $K = 2$ is able to separate samples of different tissue of origin, better than hierarchical cluster methods.



(a)



(b)

