

Figure S1: Performance evaluation on synthetic datasets. Average performance scores from 20 independent runs of each algorithm are listed. Dataset numbers correspond to the datasets listed in Table S1. K-means, hierarchical clustering, MCLUST, WGNCA and CLICK were first executed on each species and then their results were summarized across the two species. The Random clustering is generated by choosing random centroids and assigning each data point into the nearest centroid. Three performance scores were used for comparison, *average global scatter*, *average center scatter*, and proportion of genes being assigned to wrong clusters. The definitions of the first two performance scores are as follows:

1. *Average global scatter*: the average distance between every data point and the cluster center that it was assigned to. For two species, the average global scatter was the sum of distances between every gene and the center of its own species divided by the total number of genes in two species.
2. *Average center scatter*: the average distance between true cluster centers and their corresponding computed cluster centers. In a perfect clustering result, center scatter equals 0. Center scatter for two species was computed by dividing the sum of center scatters in both species by the total number of clusters in both species.

