1 Hierarchical and Admixture model comparison

In order to compare between the hierarchical and admixture models, we first draw from the entire pool of 648 Heart Left Ventricle and Muscle Skeletal samples, a set of 50 samples, some of which are Heart Left Ventricle and some are Muscle Skeletal. Now ideally our method should separate out the Heart Left Ventricle and the Muscle Skeletal samples. So, if we code Heart Left Ventricle as 1 and Muscle Skeletal as 0, then ideally after arrangement in a heatmap and once we have arranged the samples as per their clustering in the heatmap, we would expect the following ordering $t_1 = (1, 1, 1. \cdots, 1, 0, 0, 0, \cdots, 0)$ or $t_2 = (0, 0, 0, \cdots, 0, 1, 1, 1, \cdots, 1)$. These two vectors would imply complete separation of the two tissues. However, in general using hierarchical clustering or admixture, we will get some permutation of the above vectors where the number of 1 s and the number of 0 s would remain the same but the above patterns may not be maintained. Let us assume that the ordering we get by hierarchical clustering on the 50 samples be c_H and that under admixture be c_A. Then we would want to see how close these two vectors are to either of t_1. So, we define the misclassification proportion for hierarchical model and the admixture model as

$$m_H = min\left(||c_H - t_1||_0, ||c_H - t_2||_0\right)$$
$$m_A = min\left(||c_A - t_1||_0, ||c_A - t_2||_0\right)$$

Then we run 200 runs of drawing random samples of size 50 as above and obtaining m_H and m_A and then plot m_H against m_A.

The following figure depicts it for 50 samples case. I think we should repeat this with more number of samples per batch- may be 200 or 300 but it seems that Admixture is doing far better than the Hierarchical model from just these runs.

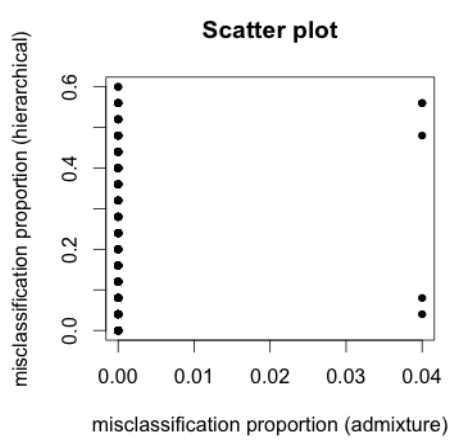

Figure 1. Comparison of the misclassification error for the Admixture model and the Hierarchical model both under average linkage and using Euclidean distance of separation.