1 Pros and Cons of Admixture

We talk about possible advantages and disadvantages that the admixture model or the topic model has over other methods like PCA, t-SNE or Hierarchical clustering.

1.1 Pros of admixture

- The Structure plot is visually easy to interpret, specially when there is an ordering among the samples and there is a continuity in the cluster patterns related to the ordering. The best example is seen in the 2010 paper by Behar et al (Figure 1)

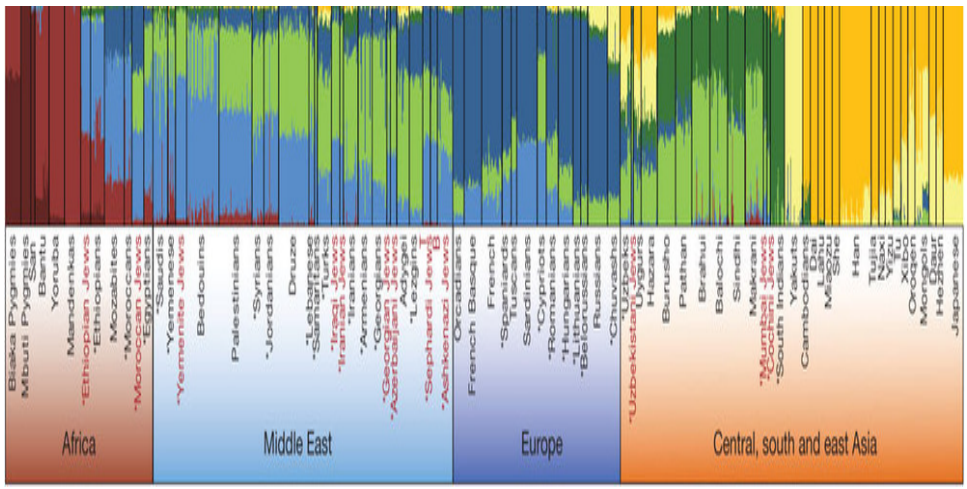

Figure 1. Structure plot in population genetics study of Jewish people using genotype data. The mixing proportions represents the proportion of admixture in modern populations grouped by continents.

We also saw a similar pattern in a single cell RNA-seq data from different parts of the renal tubule (Figure 2).

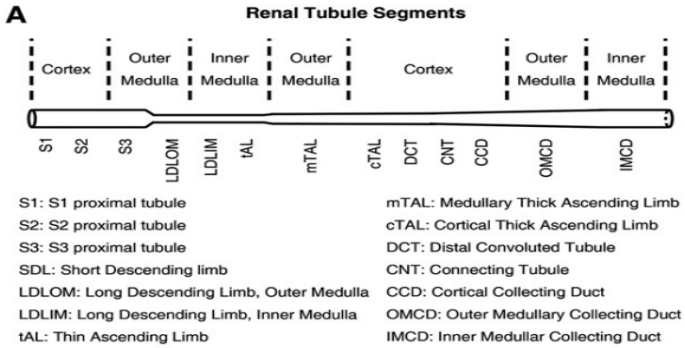

Figure 2. Renal tubule segments

The Structure plot of the cells from different parts of the renal tubule (Figure 3).

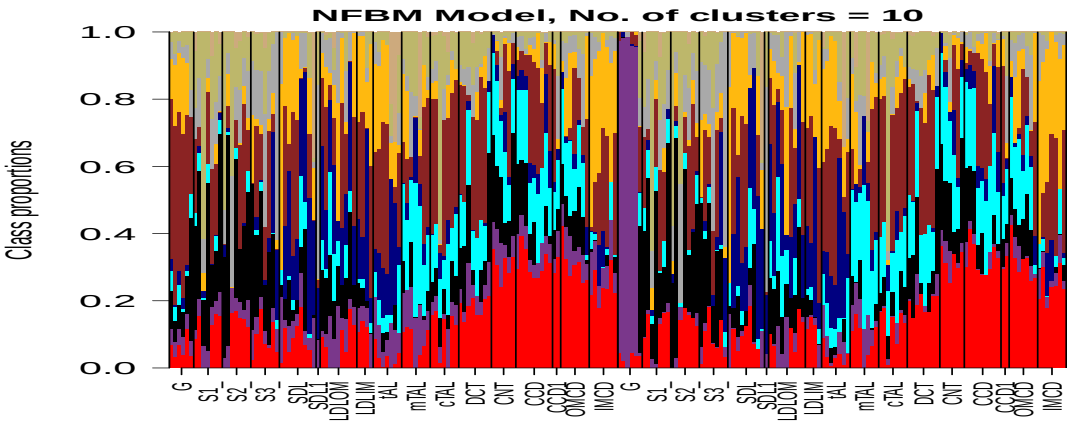

Figure 3. Structure plot of the renal tubule data

This continuity feature is pretty much exclusive to the Structure plot and is not reflected by PCa or tSNE or hierarchical clustering.

- Sometimes the interest lies in figuring out the transition point from one cluster to another and which samples represent the transition point and what are the metadata underlying those samples that could be important information. Structure plot highlights those transition points way better than PCA or t-SNE (Figure 4).

  Such a thing is difficult to figure out from PCA or t-SNE. PCA performs still better for such a data in figuring out the clusters although the transition point is difficult to obtain (Figure 5 and Figure 6).

- Structure does a better job in clustering for counts data, specially cases where there are pretty small. As you saw from the above analysis, t-SNE failed badly and even for PCA, if we did not know the cluster labels from the admixture plot, then it would have been difficult to figure out the clustering partition as it is not clear (Figure 7).

  Also when there are two groups in the dataset but they are pretty close, then Structure outperforms hierarchical clustering (figure in the RNA-seq paper).

- The topic model gives us a model loglikelihood or Bayes factor for different choices of number of clusters K. This helps us determining the optimal number of clusters and also since it is model based, it has predictive strength. If a new sample is coming in with its own set of counts, then we can predict the admixture proportions corresponding to different groups for that sample as well. Usual methods like hierarchical clustering, PCA or tSNE do not have this power.

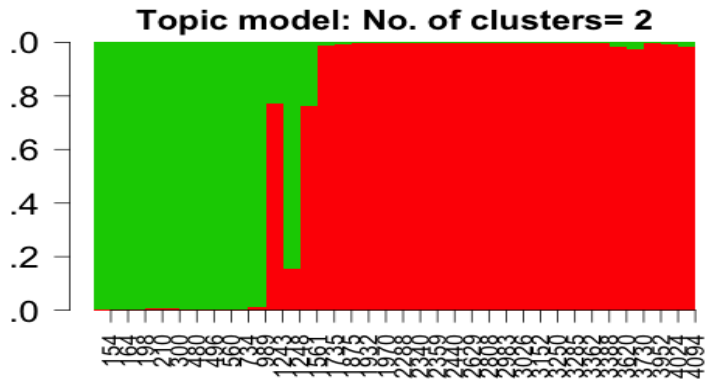

Figure 4. The Structure plot of the Himalayan forest spots based on bird abundance data with the columns representing forest spots at different elevations, the spots being arranged as per elevation. The transition from one cluster to another seems to be occurring at 1300 m approx from sea level

1.2 Cons of admixture

- The method can only be applied on counts data whereas PCA, tSNE or hierarchical clustering are general clustering tools.
- Over dispersion (mainly originating in cases of very large counts as well as very small counts) can bias the results.
- It has been observed through simulation studies that the topic proportions may not always be a reliable estimate of the actual admixture proportion of underlying subgroups. However, the list of driving OTUs or genes that drive the clusters remains preserved or is robust even when admixture proportions are off.

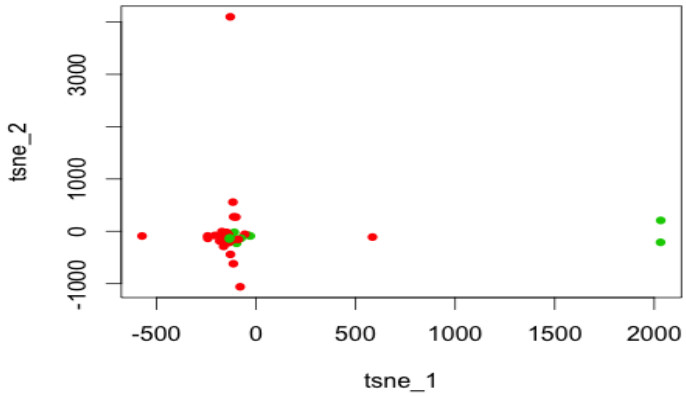

Figure 5. t-SNE plot of the forest spots.

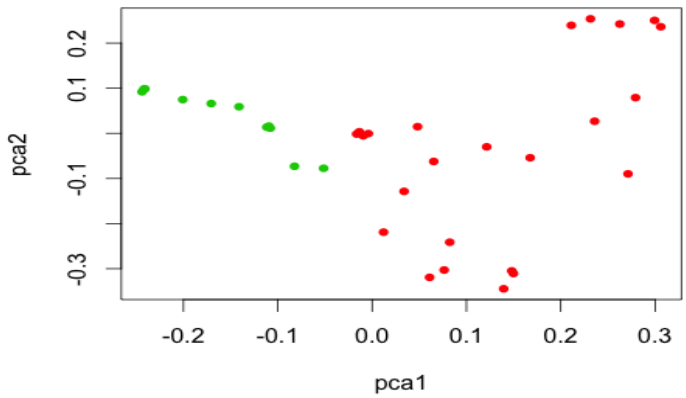

Figure 6. PCA plot of the forest spots (cluster colors obtained from admixture).

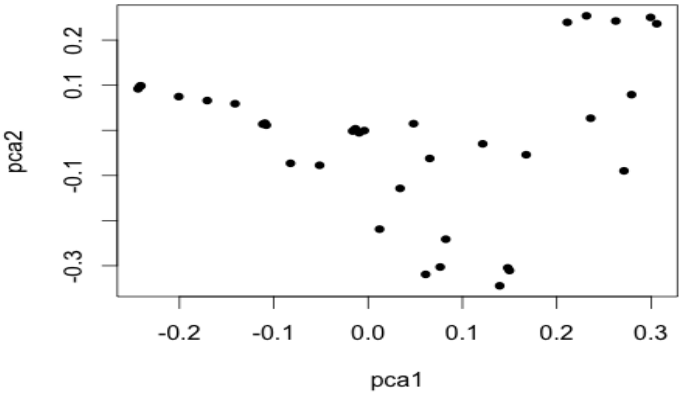

Figure 7. PCA plot of the forest spots.