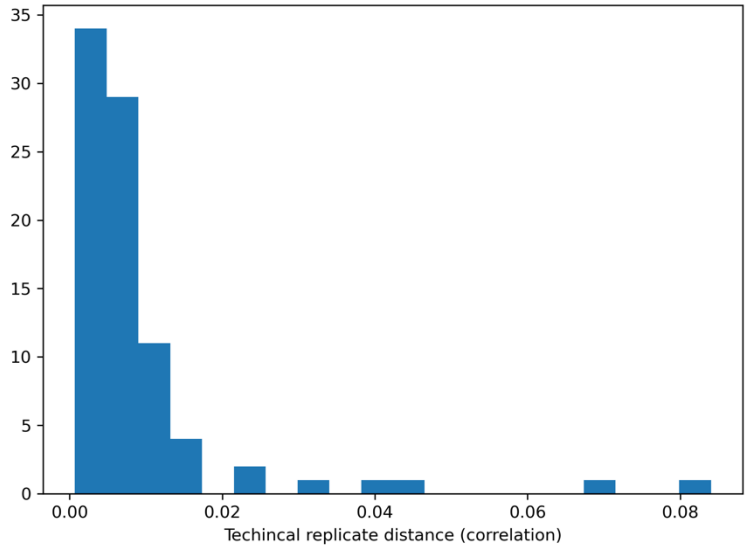
The goal of these functions is to flag two types of inconsistencies within references: divergent technical replicates and divergent biological replicates. Before running these functions, process the data through at least running clusteringDBSCAN()

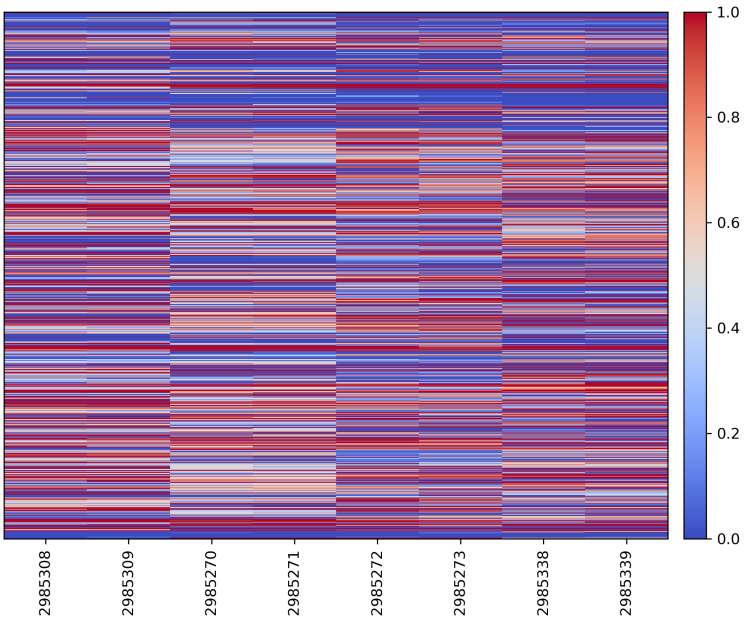
Relabeling references

For this dataset, diverging biological replicates have been renamed by adding and _a, _b, etc. to distinguish different groups. It may also be appropriate to relabel a biological replicate as a different variety if it is clear that a particular sample has been mislabeled. To entirely remove a sample from data analysis it should be renamed REMOVE. The 'references' column is the variety name that is used for all of the functions in base.py, so any reference samples with the same name in the 'references' column will be treated as the same for all of those analyses, regardless of what is in the 'reference_original' column.

Technical replicates

To generate a histogram of the distances between all the sets of technical replicates and a heatmap of the most divergent replicates run:
referenceProcessing.heatmapTechnicalRep(snpProportion, sampleMeta, distance, sample, percentile = 0.95)





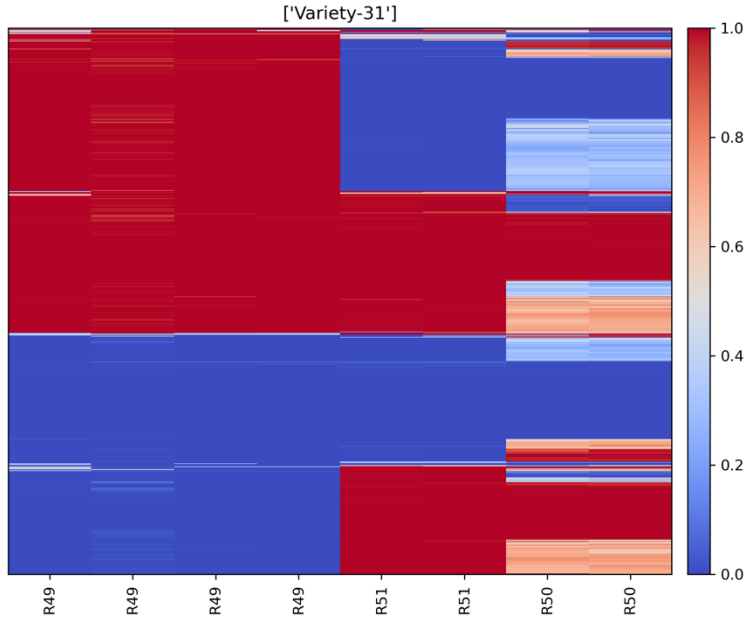
In this dataset, even the most technical replicates look reasonably similar, suggesting that none of the samples need to be removed.

Biological replicates

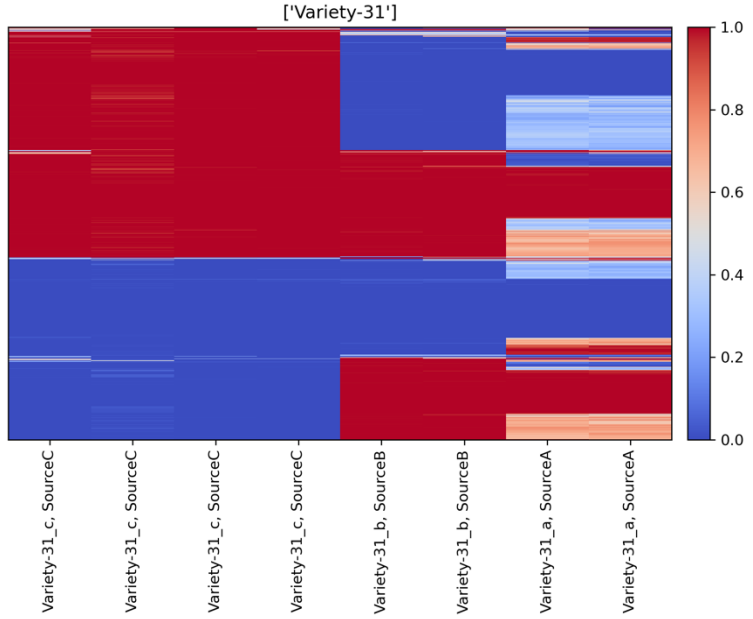
To check for differences between biological replicates run:
referenceProcessing.splitReferences(snpProportion, sampleMeta, db_communities)

The output is a list of varieties that have biological replicates in multiple clusters (note that the metadata file must have an 'inventory' column). This function not guaranteed to catch all inconsistencies within biological replicates (as is the case with Variety-31), but it is a good starting point to flag the largest differences.

For example for Variety-31, to generate a heatmap with all of biological replicates run:
plot.heatmapReferences(snpProportion, sampleMeta, ["Variety-31"], tick_type = 'inventory')

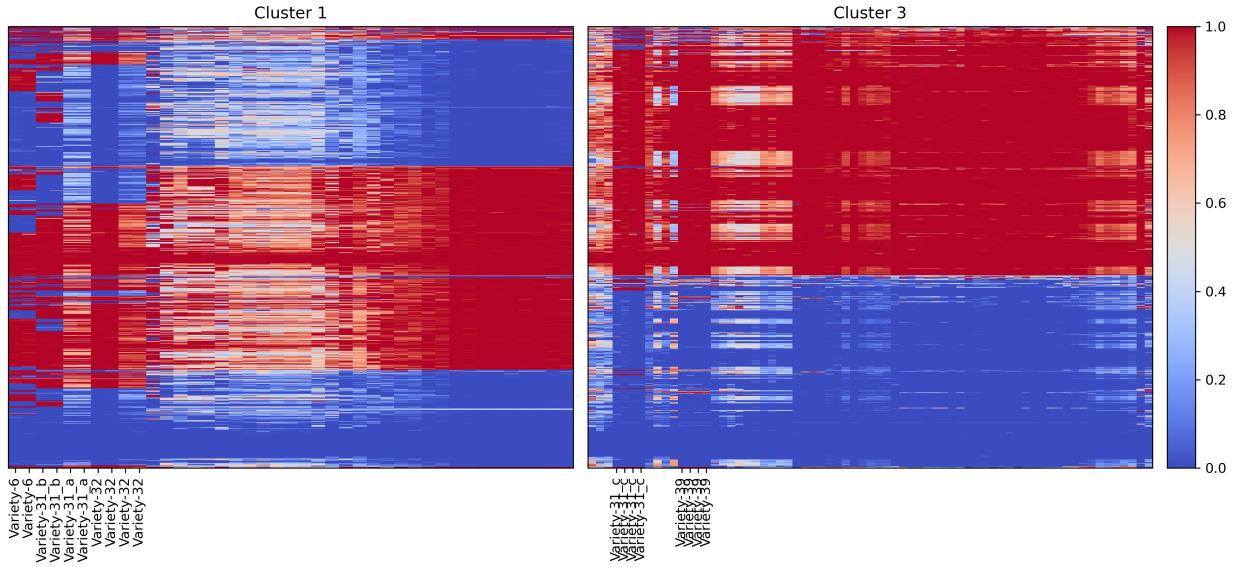


To generate the same figure but with the samples labeled by the source of the reference run:
plot.heatmapReferences(snpProportion, sampleMeta, ["Variety-31"], tick_type = 'source')



To generate a heatmap that compares the clusters that contain the biological replicates run:
plot.heatmapManyClusters(snpProportion, sampleMeta, db_communities, [1,3],

tickType='referencesAll')



For Variety-31, the technical replicates are consistent, but the three biological replicates are different. In this case, all of the biological replicates came from different sources. For a real dataset, this could be a good starting point for choosing the most authoritative biological replicate or deciding where to source additional reference material from.