# 13/SEP/19 Summary

See attached plots/data

## Multi-Genome Testing Observations

* In each Arabidopsis line, the three longer sequences have a similar ratio of matches to the final two sequences. In brassica rapa this is not true – there are very few matches for these longer substrings suggesting they are not conserved between genomes.
* Evidence for packCACTA transposons in Arabidopsis helleri from clustering.
* Count expected number of hits vs actual hits for string 5 for each Genome to investigate further.

## Clustering Observations

* Certain transposons have extremely similar r/f TIRs
* Some genomes (e.g. Arabidopsis lyrata) have clusters of great similarity – even more pronounced for the overall multi-genome plot
* Of these similar transposons, some have very similar TIRs on one end and very similar TIRs on the other, but the forward and reverse TIRs on each transposon are quite different. Correlation might be suggestive of duplication events? For example:
  + Arabidopsis helleri: 22/31, 16/41
  + Arath: 16/19, 15/21, 27/11, 25/1/12

## Algorithm Improvements

* What to blast against in the blasting step? Nt database takes too long and repbase database is inaccessible
* Use distance (dissimilarity) of r/f TIRs as a possible filtering step to remove false positives?