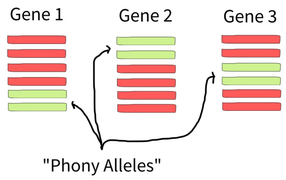
Input

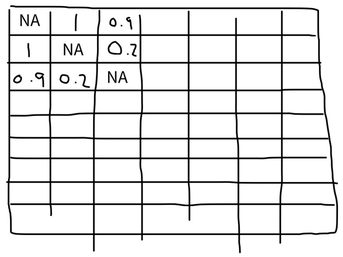
1: Locus alignments (separate concatenated exon and 'intron') containing phased allele sequences (labelled "_h1" and "h2") - any homozygous sequences are duplicated and each duplicated is arbitrarily assigned as "h1" or "h2".



Method

For each alignment:

1. Calculate pairwise distance matrix.



2. Grab the names of sample(s) that are the most similar - into table

Sample	BestMatches

BestMatches can be multiple sequences when minimum distances are equal.
Each sequence in BestMatches separated by a hyphen.

3. Remove any rows where the corresponding allele of the Sample is in BestMatches

Sample	BestMatches	
Angophora_leiocarpa_h1	Angophora_costata_costata_h2	KEEP
Angophora_bakeri_h1	Angophora_bakeri_h2-Angophora_inopina_h2	REMOVE
Angophora_mel_h1	Angophora_mel_h2	REMOVE

4. Record the names from BestMatches for each Sample.

e.g. Angophora_leiocarpa_h1 Angophora_costata_costata_h2

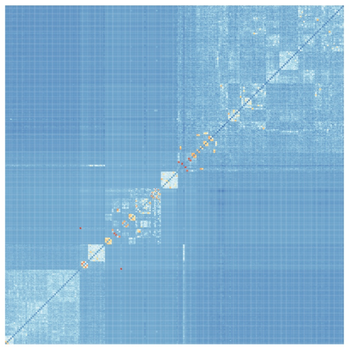
This done for each alignment so you end up with a tally, like:

Gene 1	Angophora_leiocarpa_h1	Angophora_costata_costata_h2
Gene 2	Angophora_leiocarpa_h1	Angophora_costata_costata_h2, Angophora_bakeri_h1
Gene 3	← Blank for genes where the alleles are closest pairs	
Gene 4	Angophora_leiocarpa_h1	Angophora_costata_costata_h2, Angophora_mel_h1
Gene 5	Angophora_leiocarpa_h1	Angophora_costata_costata_h2

5. Tally up the number of times each allele is paired in a frequency table.

	Angophora_leiocarpa_h2	A_costata_h2	A_bakeri_h1	A_mel_h1
Angophora_leiocarpa_h1	0	4	1	1

6. Plot this as a heat map



7. Count the number of times each Sample is split

Angophora_leiocarpa_h1 n = 4 + 1 + 1 = 6

8. Table with this info

Sample	n
RMF503_Corymbia_abergiana_h1	705
RMF564_Corymbia_stockeri_stockeri_h1	693
RMF564_Corymbia_stockeri_stockeri_h2	686
RMF503_Corymbia_abergiana_h2	666
RMF524_Corymbia_lamprophylla_h1	655
RMF524_Corymbia_lamprophylla_h2	651
RMF536_Corymbia_serendipita_h2	651
RMF536_Corymbia_serendipita_h1	647
MJB2525_Corymbia_abbreviata_h2	635
RMF520_Corymbia_ellipsoidea_h1	635
MJB2525_Corymbia_abbreviata_h1	634
RMF520_Corymbia_ellipsoidea_h2	631
RMF505_Corymbia_ellipsoidea_h1	624
RMF577_Corymbia_disjuncta_h2	624
RMF505_Corymbia_ellipsoidea_h2	622

9. Filter to include samples with highest n

By default I've set this to the max number of samples in any alignment divided by 2

10. For each sequence passing this filter, get the name of its highest-pairing sequence across all genes.

These get associated in a one-to-one renaming table.

sample	name2match
CANB407099_Corymbia_lenziana_h1	CANB5135651_Corymbia_eremaea
CANB407099_Corymbia_lenziana_h2	CANB5135651_Corymbia_eremaea
CANB5135651_Corymbia_eremaea_h1	CANB407099_Corymbia_lenziana
CANB5135651_Corymbia_eremaea_h2	CANB407099_Corymbia_lenziana
CANB6444091_Corymbia_torta_h1	MJB2511_Corymbia_grandifolia
CANB6444091_Corymbia_torta_h2	PDE10_Corymbia_dendromerinx
CANB7311121_Corymbia_foelscheana_h1	RMF549_Corymbia_greeniana
CANB7311121_Corymbia_foelscheana_h2	MJB2540_Corymbia_foelscheana
CANB817975_Corymbia_kombolgiensis_h1	MJB2544_Corymbia_polysciada

11. Sorting Alleles

11a Re-run steps 1 and 2: To calculate pairwise distances between each sample in each alignment and get the closest allele for each sample

11b For each sample in each alignment, if the sample does not go with its corresponding allele, and if it does go with its highest pair across all genes (i.e. the name in 'name2match' in Step 10), then rename that Sample to 'Sample_matched_to_name2match'.

11c Write out the alignment as a fasta