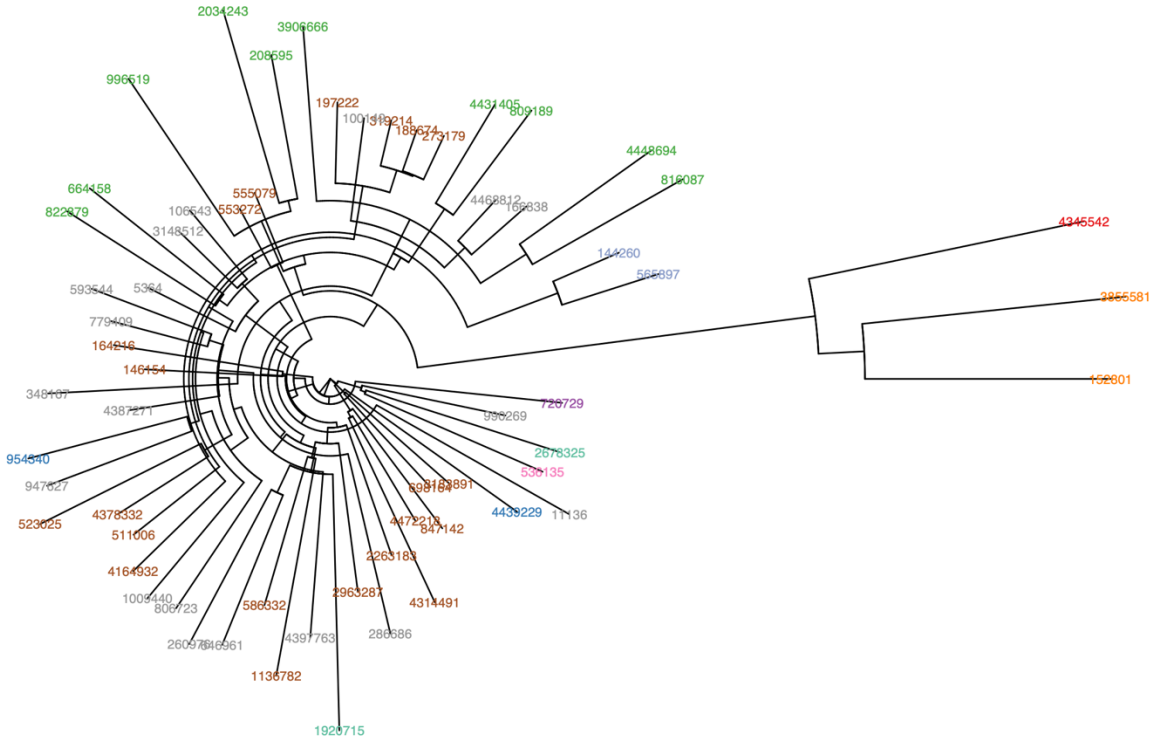
CSCI 5481
Nicole Sullivan

Question 3

Find a way to visualize your trees from step (3). If you made the output correctly for step (3), you can use the included R script (after installing the “ape” and the “RColorBrewer” package as described above).



Question 4

Why are tips of similar color mostly clustered together?

The colors indicate the different phyla of the sequences, and *usually* (though not always) taxonomic classifications group organisms with similar sequences into the same taxonomical unit, in this case phyla.

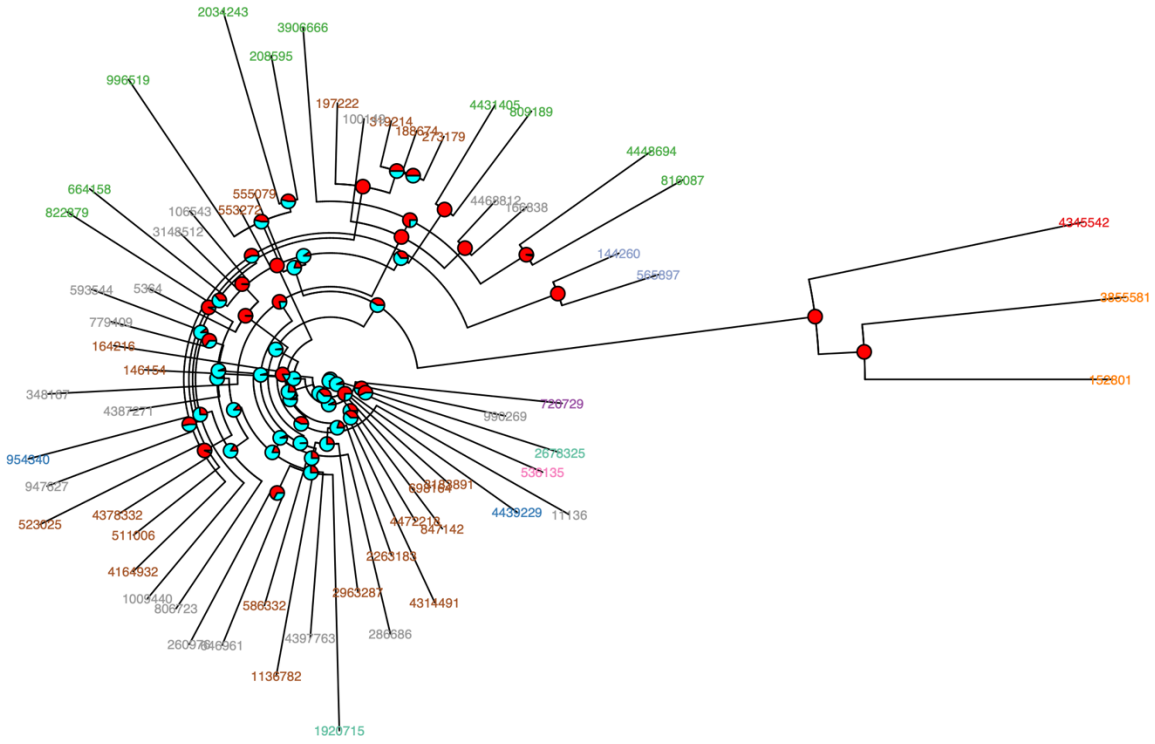
Question 5

Name 2 distinct reasons why the clustering by color is not perfect.

I think this could be for a few different reasons. One reason could be because the similarity calculation is dependent on the penalties used with the input multiple sequence alignment – in this implementation of Nei-Saitou, a gap vs. a base in another sequence will count towards the dissimilarity between those sequences. So the MSA may have allowed for many gaps to be inserted for reasons known via domain knowledge, causing the sequences to have low similarity when using such a coarse metric – but in reality, they may have high similarity amongst the aligned (non-gap) bases.

Another reason could be how phyla classifications have historically been made. As far as I understand it, the earliest taxonomic classifications were based on phenotype only; and later, when the microscope was invented, microscopic examination of organisms heavily influenced the organism's classification. However, sequencing is revealing that some organisms that once appeared very different by these methods are actually not very different at all at the sequence-level. In the Bracken paper, we read that different taxonomic classifications persist today for organisms that have actually been resolved via sequence as the same. For example, historically *M. bovis and M. tuberculosis* have been classified as separate species, but the paper says that "high sequence identity indicates that they should be considered as two strains of a single species" (p. 3). So some of the differences may arise from discrepancies in sequence-based classification and microscopic- or phenotypic-based classification, which is where many of our taxonomic classifications came from.

Question 7



Question 8

Based on the bootstrap tree pictured above, why is bootstrap support generally higher for the internal nodes closes to the tips, and lower for the internal nodes closest to the root?

Joins closer to the tips generally occur first (since at first there are no, or very few, internal nodes available to join). Because of the granularity at the tip-level, it's likely that there are very few equivalent joins across the tips, if any, so there's less variation in joining here, corresponding to higher support. However, as more joins occur, the distance between nodes becomes smaller, and therefore there are many

more possible equivalent joins. This high variation in possibilities means lower support for any one given join at that level.

Put another way (in terms of the biological information encoded), starting from the sequences we have today, we have less confidence the further back in time we try to infer splits. At the tips of the tree, there's high resolution (sequence base pairs), whereas near the root of the tree (i.e. closer to the common ancestor), the resolution is lower – the divergence events to which these higher nodes correspond occurred in the ancient past, and there was higher sequence similarity between those organisms (they weren't differentiated into the taxonomies we see today). So our bootstrap support degrades as we get farther from our original sequences.