Supplementary Material

Appendix A

To optimize the choice of r and b, we used 100 random samples of 10,000 sequences from the COI database from the CALeDNA project (Curd *et al.*, 2019) and first chose the values of 15, 100, 250, 500, 750, and 1000 for r and set b to 15. We calculated *relative NMI*, *relative purity*, and *relative incompatibility* as described in the Methods. *Relative NMI* (Figure S17) and *relative purity* (Figure S18) tend to increase with increasing values of r for species, genus, and family taxonomic levels. Additionally, *relative incompatibility* (Figure S19) tends to decrease with increasing values of r for species, genus, family, order, and class taxonomic levels. We chose the value of r to be 750 since it maximized all three of the performance measures at species, genus, and family taxonomic levels. To optimize b, we set r to 750 and chose values of 5, 10, 15, 20, 50, and 80 for b for the same dataset. We also calculated *relative NMI* (Figure S20), *relative purity* (Figure S21), and *relative incompatibility* (Figure S22). While the choice of b on *relative NMI* (Figure S20) and *relative purity* (Figure S21) showed opposite effects, the choice of b had little effect on *relative incompatibility* (Figure S22). *Relative NMI* tends to increase as b becomes smaller but *relative purity* tends to decrease as b becomes smaller. We chose the value of b to be 15 which maximized *relative incompatibility* at the species, genus, and family taxonomic levels. While this procedure for choosing r and b is based on a specific data set of COI sequences, we do not observe great dependence of the performance on the exact values of r and b (see Results) and recommend them for use in analyses of other data sets as well in the absence of other information.

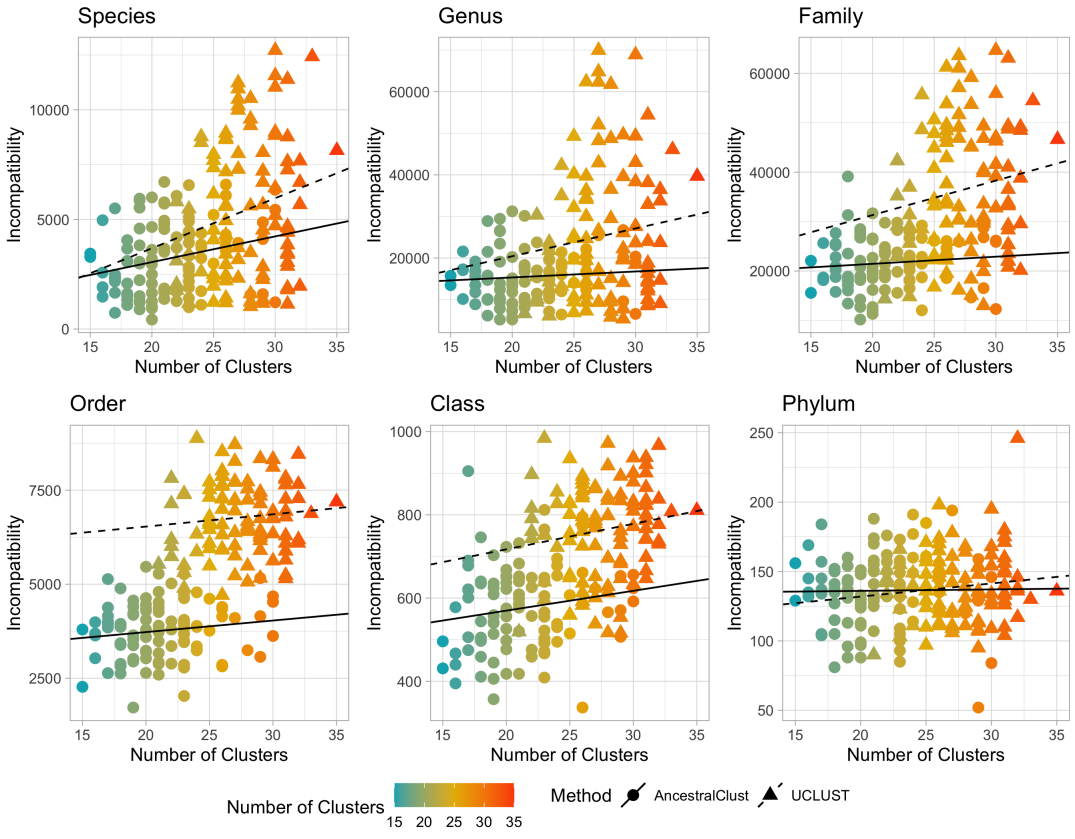

Figure S1. Linear regression between incompatibility and number of clusters using AncestralClust (solid line) and UCLUST (dotted line) for 100 samples of 10,000 randomly chosen COI reference sequences from the CALeDNA Project (Curd *et al.*, 2019).

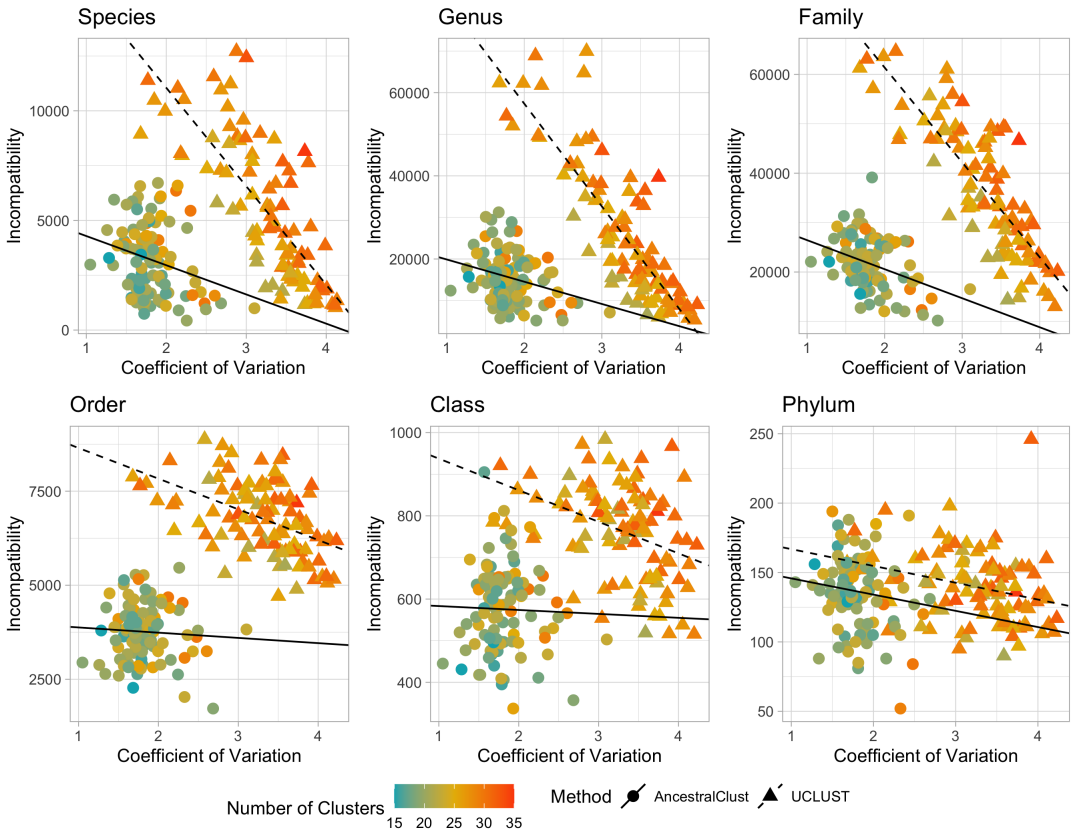

Figure S2. Linear regression between incompatibility and Coefficient of Variation using AncestralClust (solid line) and UCLUST (dotted line) for 100 samples of 10,000 randomly chosen COI reference sequences from the CALeDNA Project (Curd *et al.*, 2019).

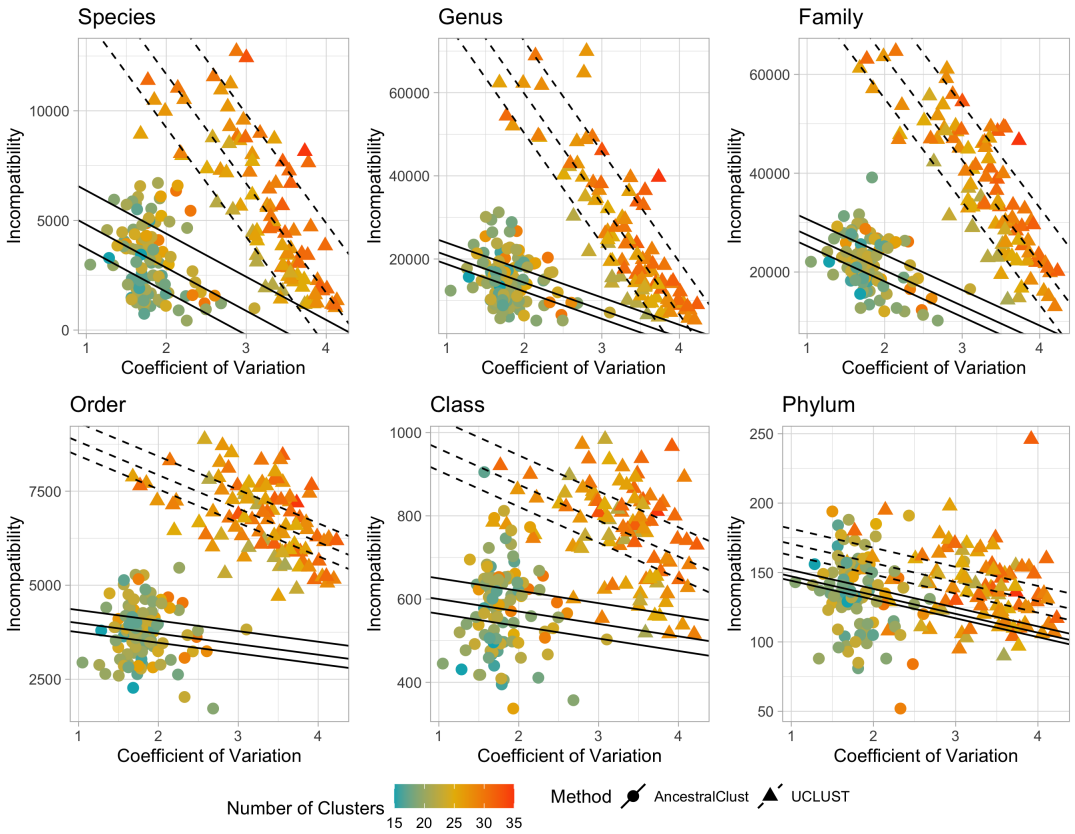

Figure S3. Multiple regression between incompatibility and number of clusters and Coefficient of Variation using AncestralClust (solid line) and UCLUST (dotted line) for 100 samples of 10,000 randomly chosen COI reference sequences from the CALeDNA Project (Curd *et al.*, 2019). The maximum, mean, and minimum number of clusters for the respective methods are shown.

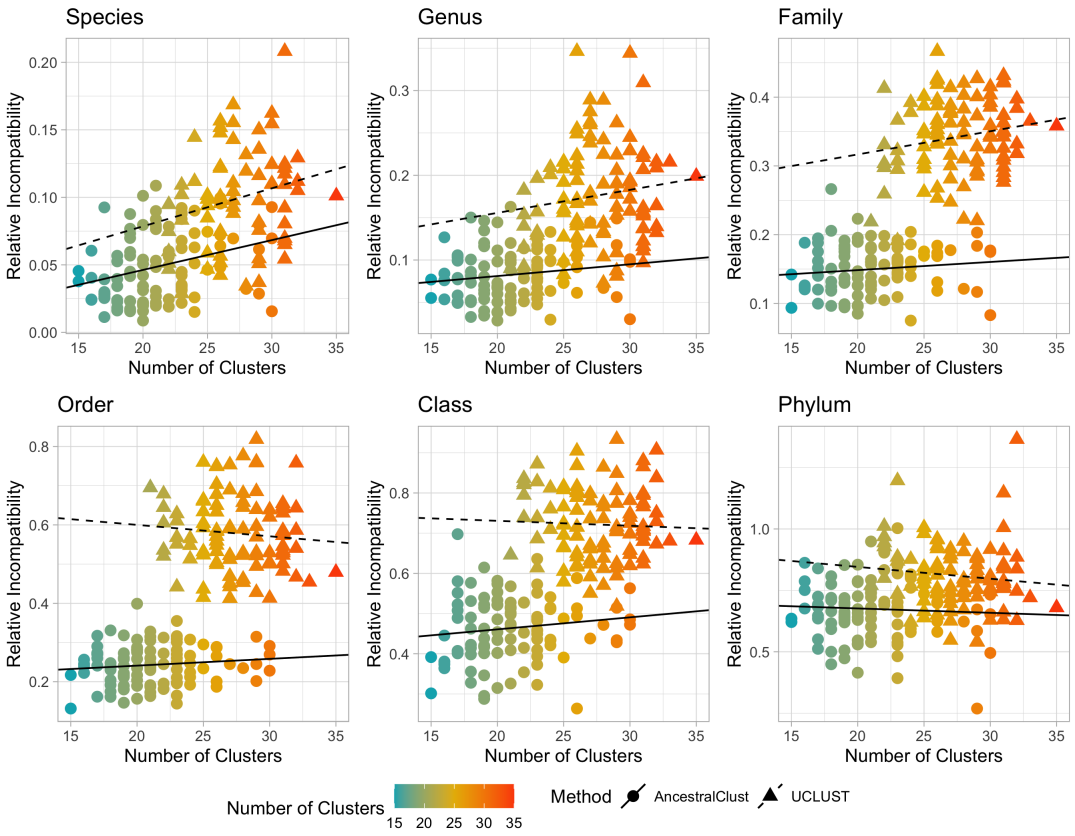

Figure S4. Linear regression between *relative incompatibility* and number of clusters using AncestralClust (solid line) and UCLUST (dotted line) for 100 samples of 10,000 randomly chosen COI reference sequences from the CALeDNA Project (Curd *et al.*, 2019).

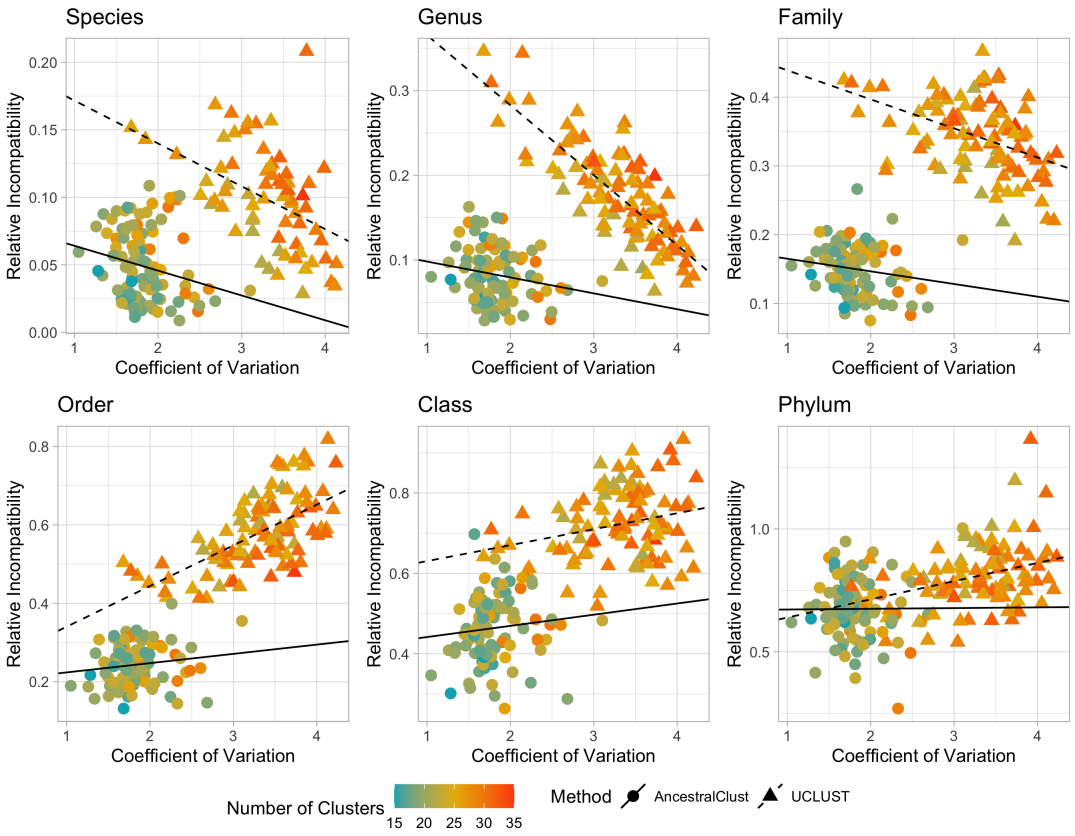

Figure S5. Linear regression between *relative incompatibility* and Coefficient of Variation using AncestralClust (solid line) and UCLUST (dotted line) for 100 samples of 10,000 randomly chosen COI reference sequences from the CALeDNA Project (Curd *et al.*, 2019).

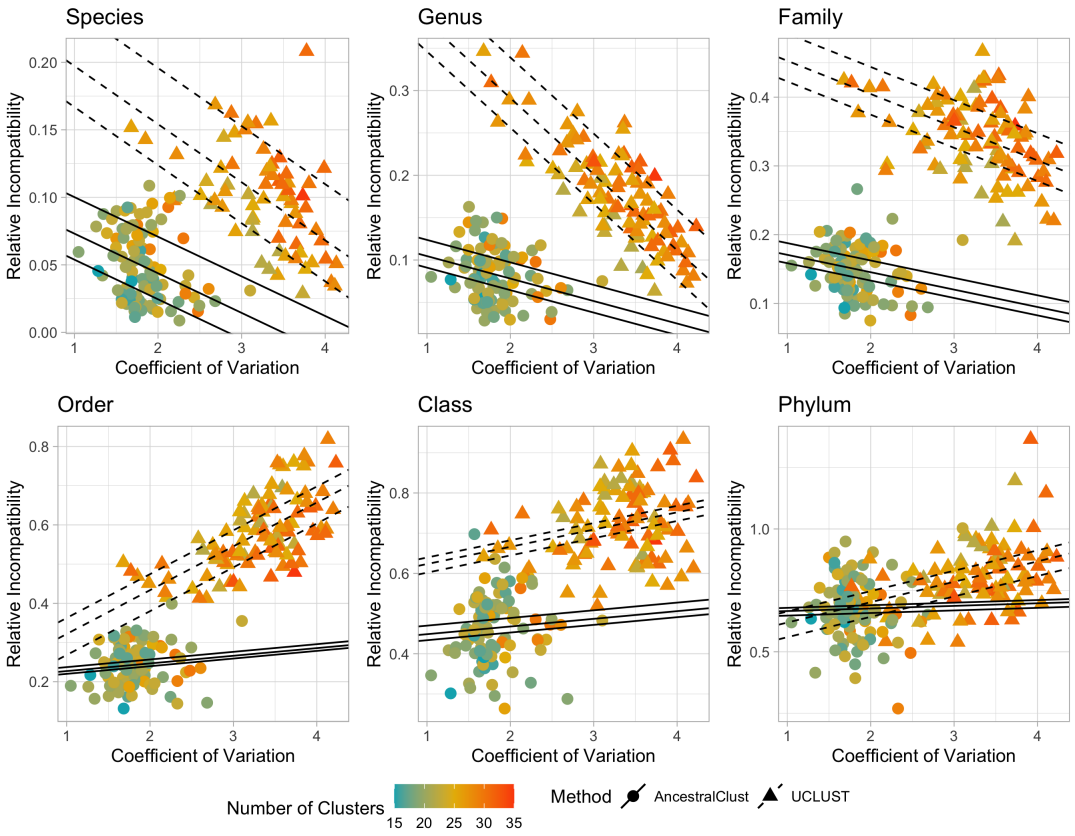

Figure S6. Multiple regression between *relative incompatibility* and number of clusters and Coefficient of Variation using AncestralClust (solid line) and UCLUST (dotted line) for 100 samples of 10,000 randomly chosen COI reference sequences from the CALeDNA Project (Curd *et al*., 2019). The maximum, mean, and minimum number of clusters for the respective methods are shown.

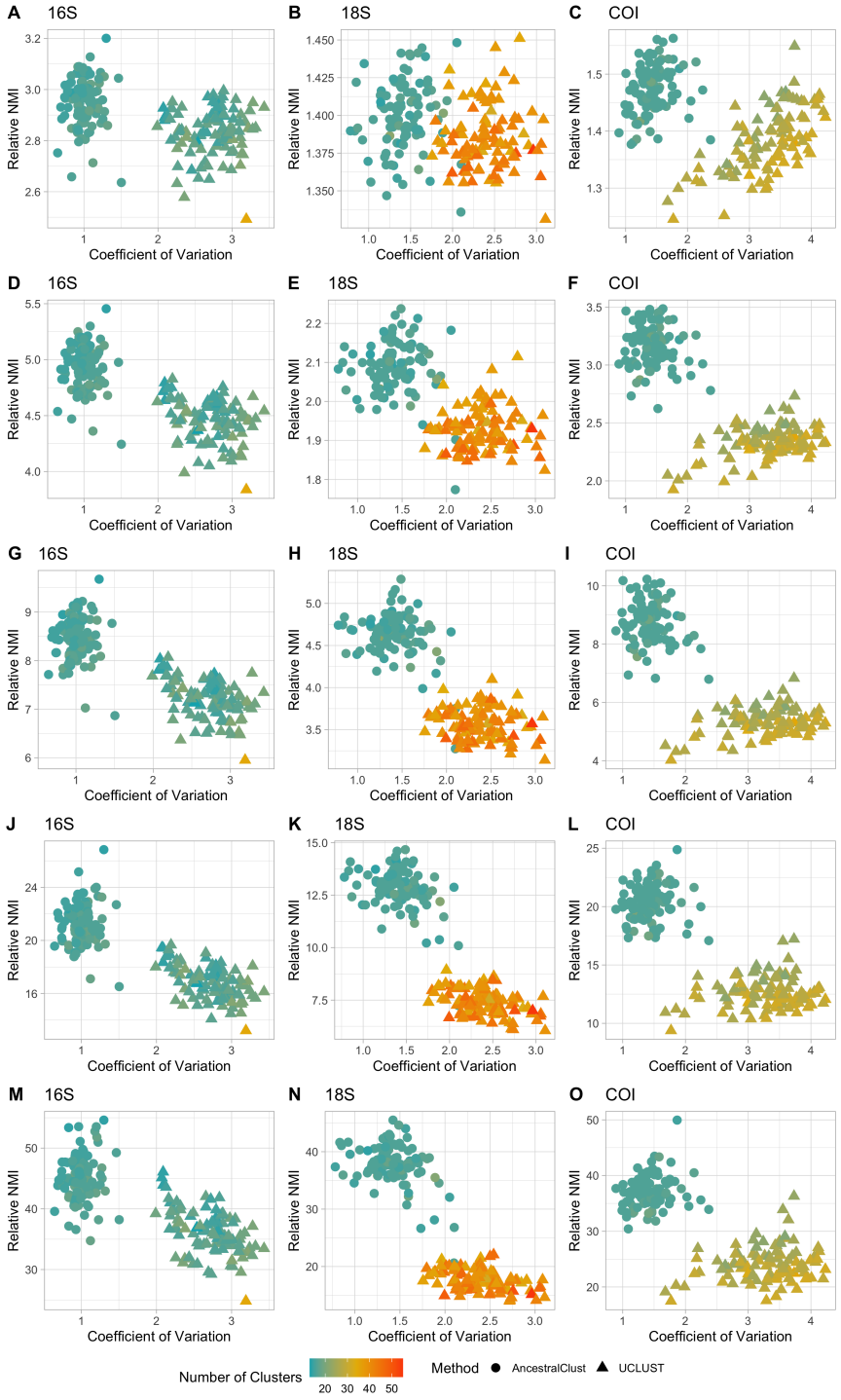

Figure S7. *Relative NMI* against Coefficient of Variation for AncestralClust and UCLUST for 100 samples of 10,000 randomly chosen 16S, 18S, and COI reference sequences for taxonomic levels genus (A-C), family (D-F), order (G-I), class (J-L), and phylum (M-O). All reference sequences are from the CALeDNA Project (Curd *et al.*, 2019). The similarity threshold for UCLUST for 16S and 18S is 0.58, and for COI the similarity threshold is 0.62. For AncestralClust, we used 750 initial random sequences with 15 initial clusters.

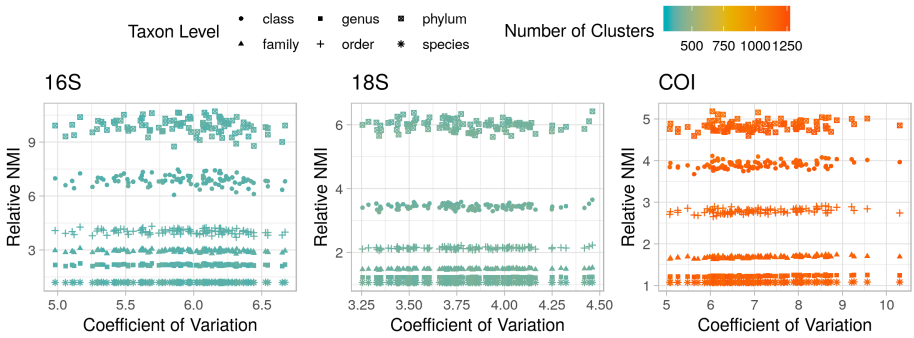

Figure S8. *Relative NMI* at all taxon levels for CD-HIT against coefficient of variation for 100 samples of 10,000 randomly chosen 16S, 18S, and COI reference sequences from the CALeDNA Project (Curd *et al.*, 2019). The similarity threshold for CD-HIT is 0.8.

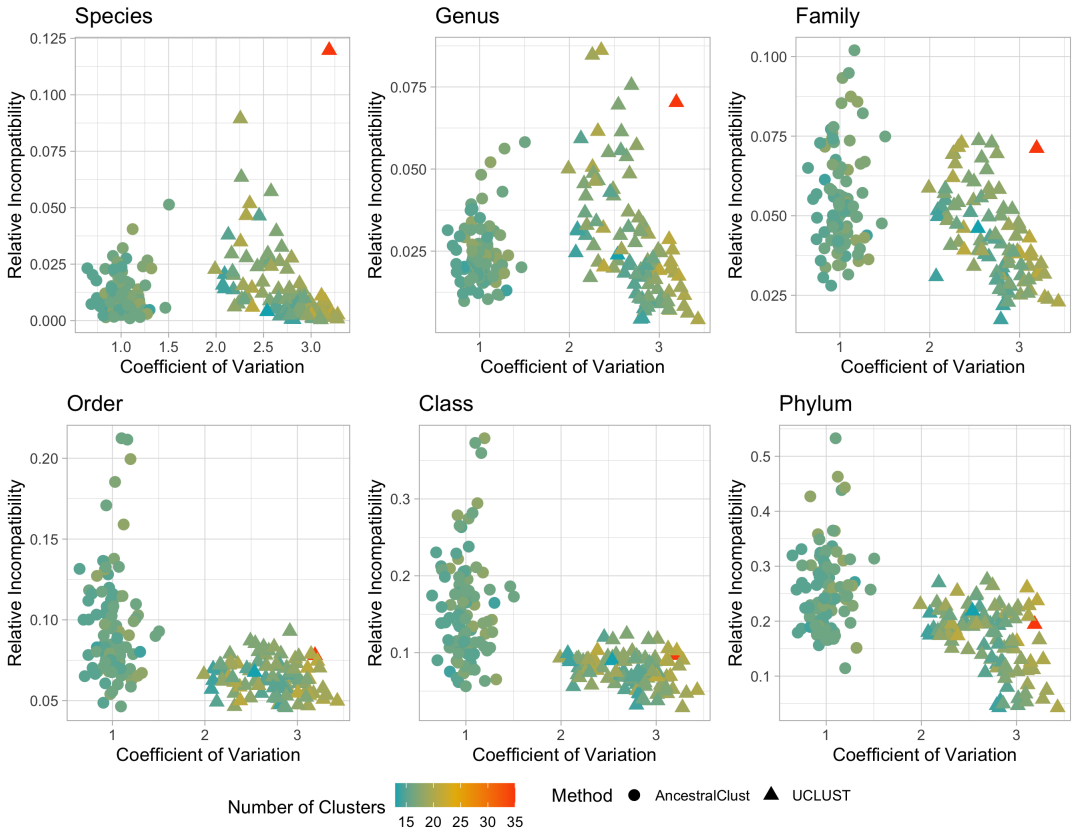

Figure S9. Relative incompatibility against Coefficient of Variation for AncestralClust and UCLUST for 100 samples of 10,000 randomly chosen 16S reference sequences. 16S reference sequences are from the CALeDNA Project (Curd *et al.*, 2019). The similarity threshold for UCLUST was 0.58. For AncestralClust, we used 750 initial random sequences with 15 initial clusters.

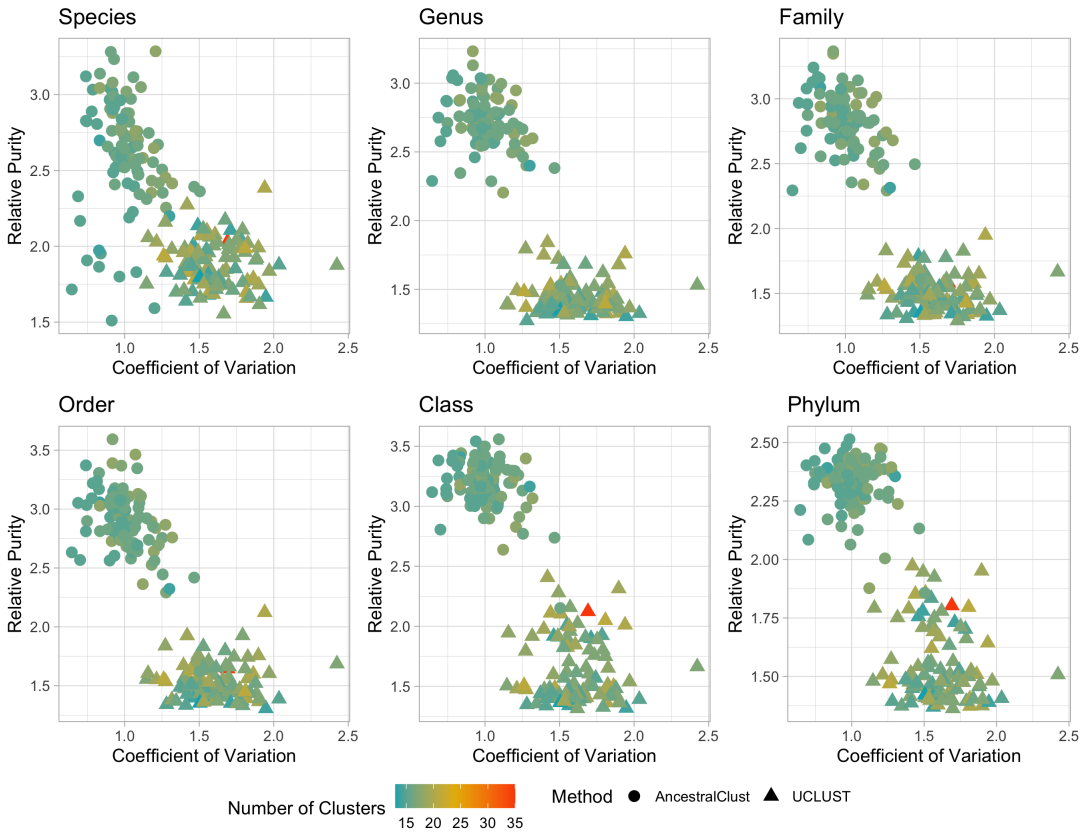

Figure S10. *Relative purity* against Coefficient of Variation for AncestralClust and UCLUST for 100 samples of 10,000 randomly chosen 16S reference sequences. 16S reference sequences are from the CALeDNA Project (Curd *et al.*, 2019). The similarity threshold for UCLUST was 0.58. For AncestralClust, we used 750 initial random sequences with 15 initial clusters.

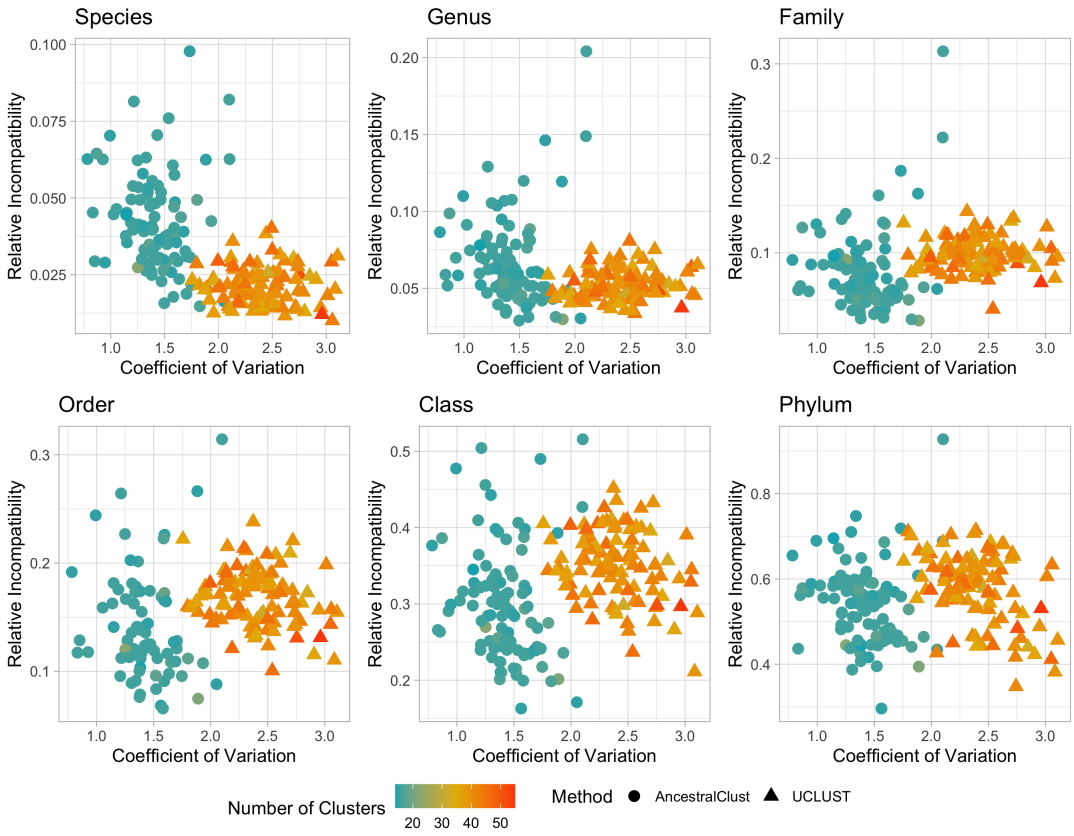

Figure S11. *Relative incompatibility* against Coefficient of Variation for AncestralClust and UCLUST for 100 samples of 10,000 randomly chosen 18S reference sequences. 18S reference sequences are from the CALeDNA Project (Curd *et al.*, 2019). The similarity threshold for UCLUST was 0.58. For AncestralClust, we used 750 initial random sequences with 15 initial clusters.

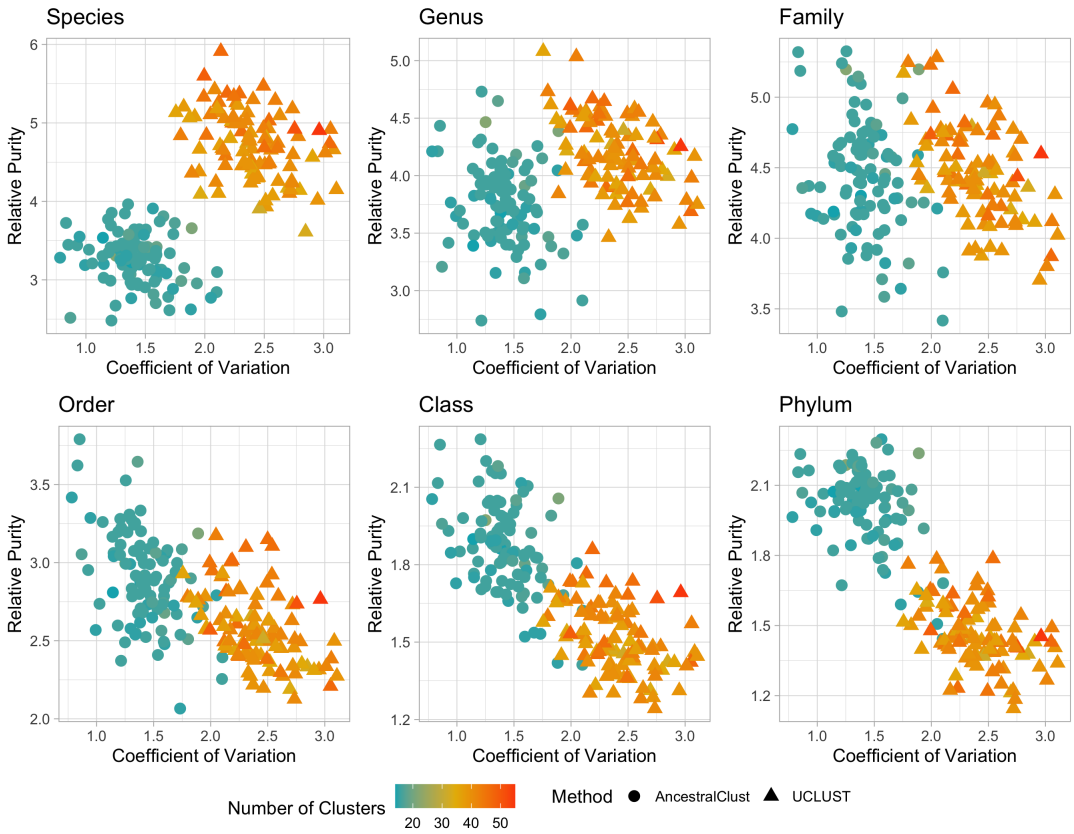

Figure S12. *Relative purity* against Coefficient of Variation for AncestralClust and UCLUST for 100 samples of 10,000 randomly chosen 18S reference sequences. 18S reference sequences are from the CALeDNA Project (Curd *et al.*, 2019). The similarity threshold for UCLUST was 0.58. For AncestralClust, we used 750 initial random sequences with 15 initial clusters.

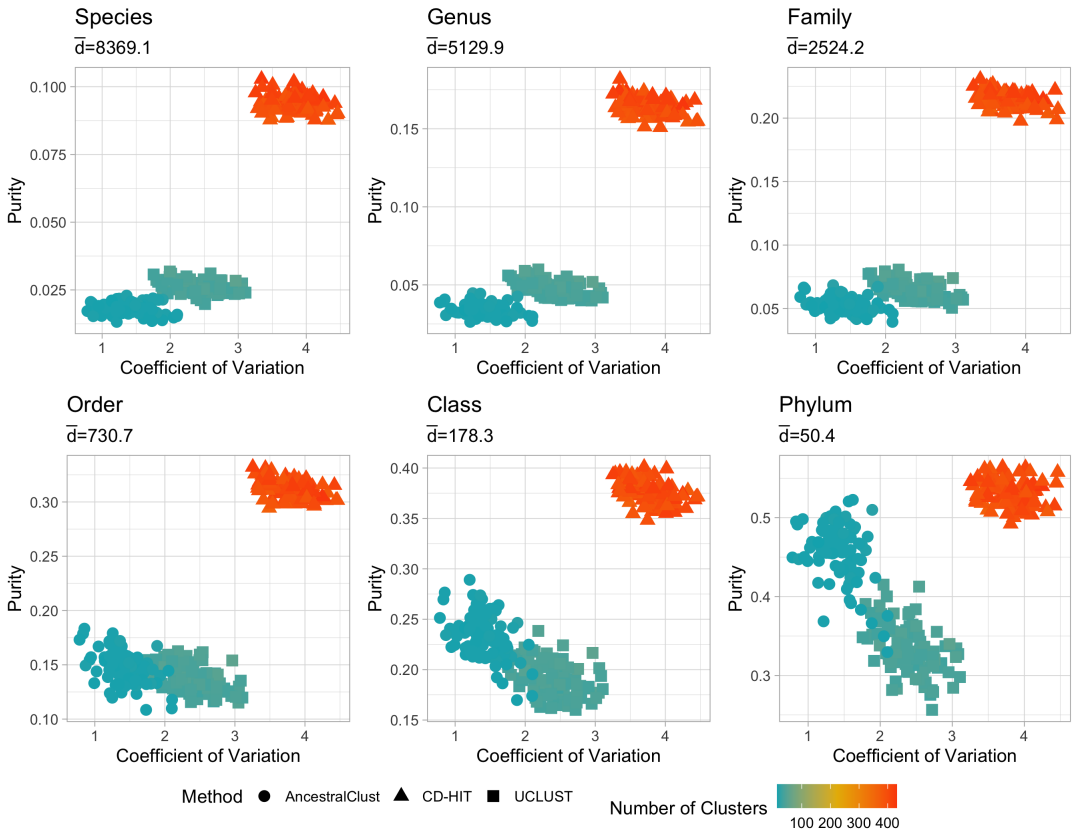

Figure S13. Purity against Coefficient of Variation for AncestralClust, UCLUST, and CD-HIT for 100 samples of 10,000 randomly chosen 18S reference sequences. 18S reference sequences are from the CALeDNA Project (Curd *et al.*, 2019). The similarity threshold for UCLUST was 0.58 and for CD-HIT was 0.8. For AncestralClust, we used 750 initial random sequences with 15 initial clusters. $\bar{d}$ is the average number of taxonomic groups over 100 samples of each taxonomic level.

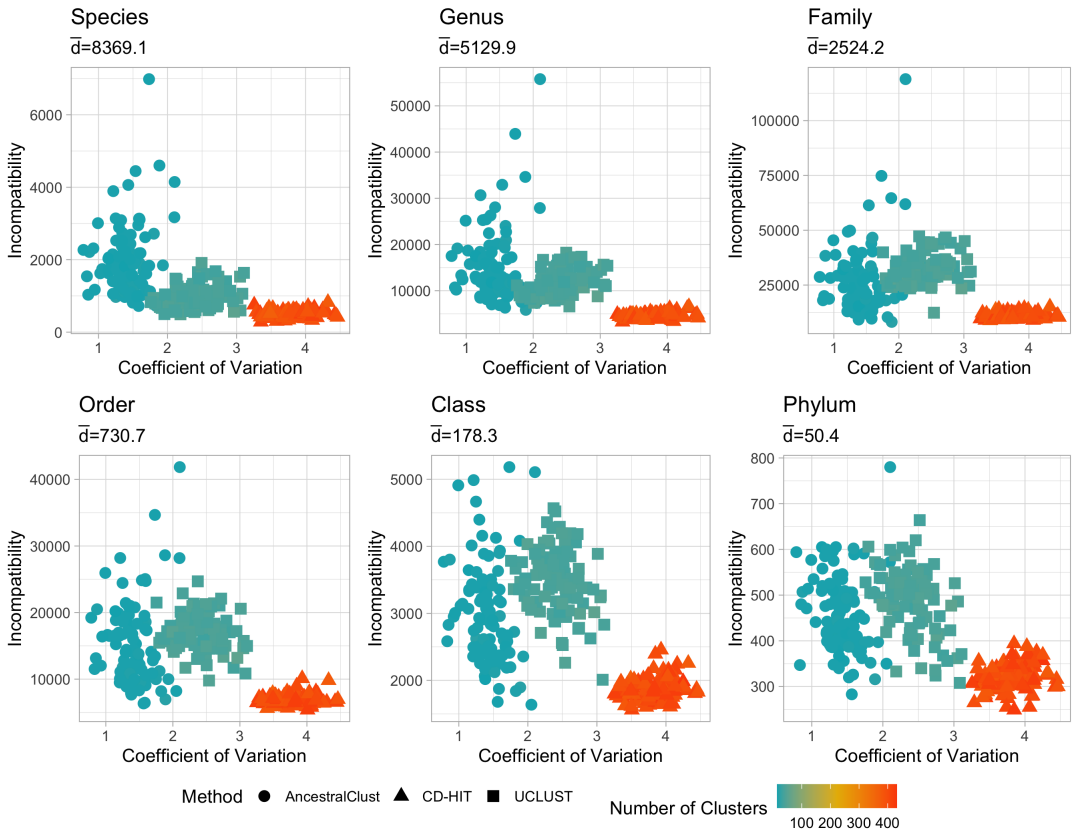

Figure S14. Incompatibility against Coefficient of Variation for AncestralClust, UCLUST, and CD-HIT for 100 samples of 10,000 randomly chosen 18S reference sequences. 18S reference sequences are from the CALeDNA Project (Curd *et al.*, 2019). The similarity threshold for UCLUST was 0.58 and for CD-HIT was 0.8. For AncestralClust, we used 750 initial random sequences with 15 initial clusters. $\bar{d}$ is the average number of taxonomic groups over 100 samples of each taxonomic level.

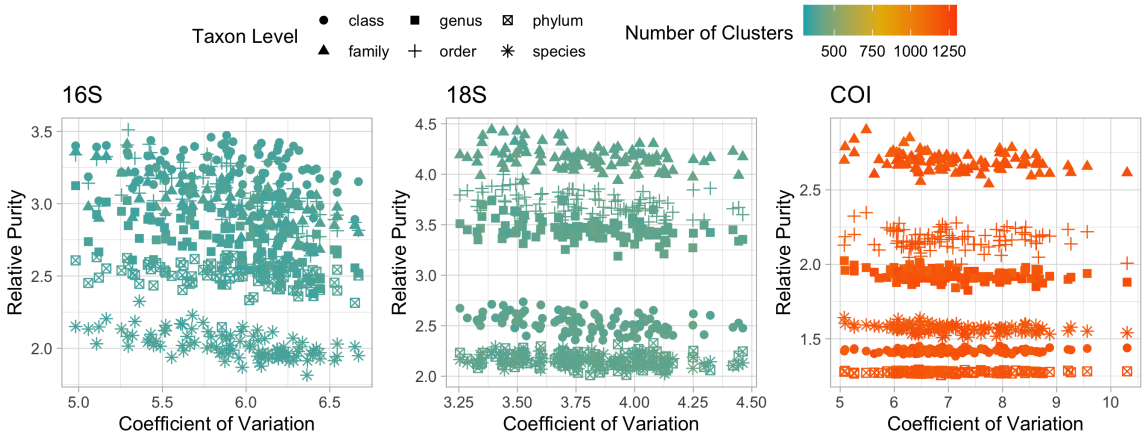

Figure S15. *Relative purity* at all taxon levels for CD-HIT against Coefficient of Variation for 100 samples of 10,000 randomly chosen 16S, 18S, and COI reference sequences from the CALeDNA Project (Curd *et al.*, 2019). The similarity threshold for CD-HIT is 0.8.

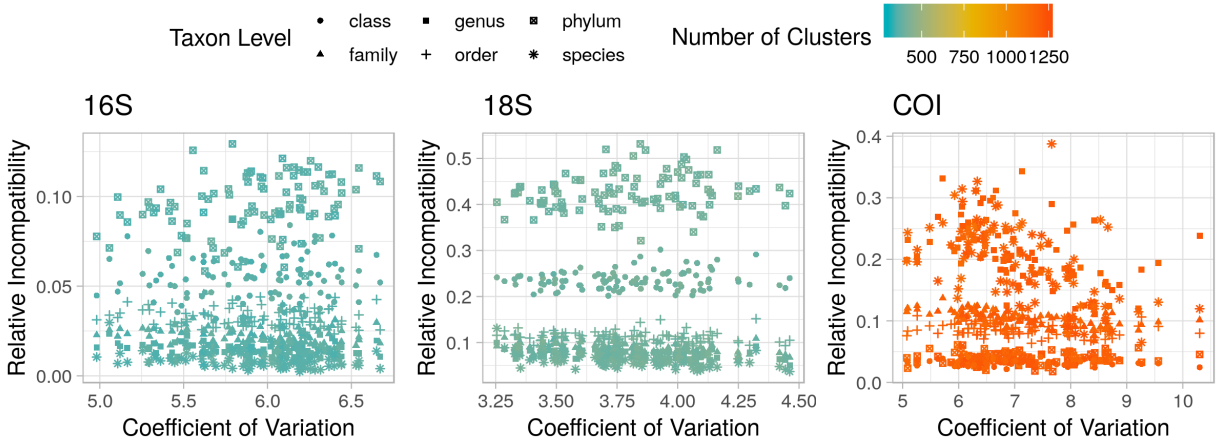

Figure S16. *Relative incompatibility* at all taxon levels for CD-HIT against Coefficient of Variation for 100 samples of 10,000 randomly chosen 16S, 18S, and COI reference sequences from the CALeDNA Project (Curd *et al.*, 2019). The similarity threshold for CD-HIT is 0.8.

Table S1. Taxonomy of sequences used for analysis in Table 1.

Taxonomy
Eukaryota;Apicomplexa;Aconoidasida;Haemosporida;Plasmodiidae;Plasmodium;Plasmodium vivax
Eukaryota;Arthropoda;Arachnida;Araneae;Araneidae;Argiope;Argiope bruennichi
Eukaryota;Arthropoda;Collembola;NA;Entomobryidae;Entomobrya;Entomobrya sp. BOLD:ACL6239
Eukaryota;Arthropoda;Insecta;Hymenoptera;Diapriidae;NA;Diapriidae sp. BOLD-2016
Eukaryota;Arthropoda;Maxillopoda;Sessilia;Chthamalidae;Notochthamalus;Notochthamalus scabrosus
Eukaryota;Chordata;Mammalia;Carnivora;Canidae;Canis;Canis lupus
Eukaryota;Echinodermata;Asteroidea;Valvatida;Ophidiasteridae;Linckia;Linckia laevigata
Eukaryota;Mollusca;Bivalvia;Mytiloida;Mytilidae;Mytilus;Mytilus trossulus
Eukaryota;Mollusca;Gastropoda;NA;Littorinidae;Melarhaphe;Melarhaphe neritoides
Eukaryota;Nematoda;Chromadorea;Rhabditida;Strongyloididae;Strongyloides;Strongyloides stercoralis
Eukaryota;Platyhelminthes;Cestoda;Diphyllobothriidea;Diphyllobothriidae;Schistocephalus;Schistocephalus solidus

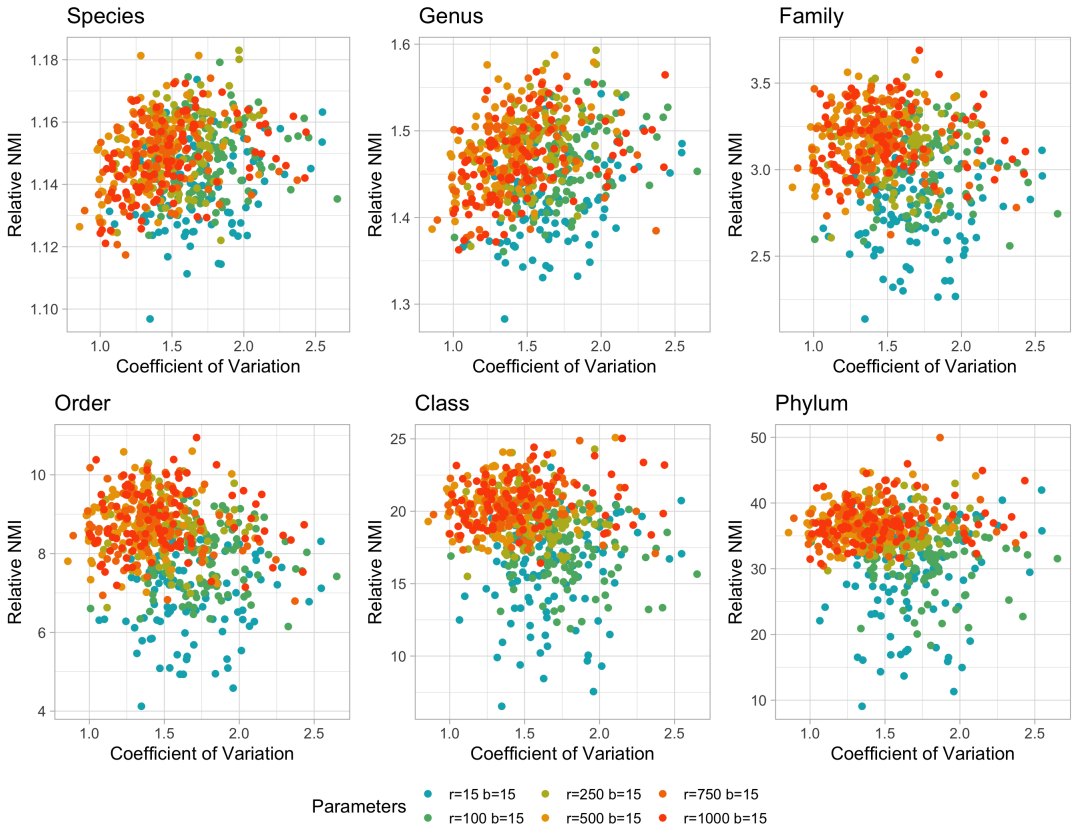

Figure S17. *Relative NMI* against Coefficient of Variation for differing values of r ($r = 15$, $r = 100$, $r = 250$, $r = 500$, $r = 750$, and $r = 1000$) with $b = 15$.

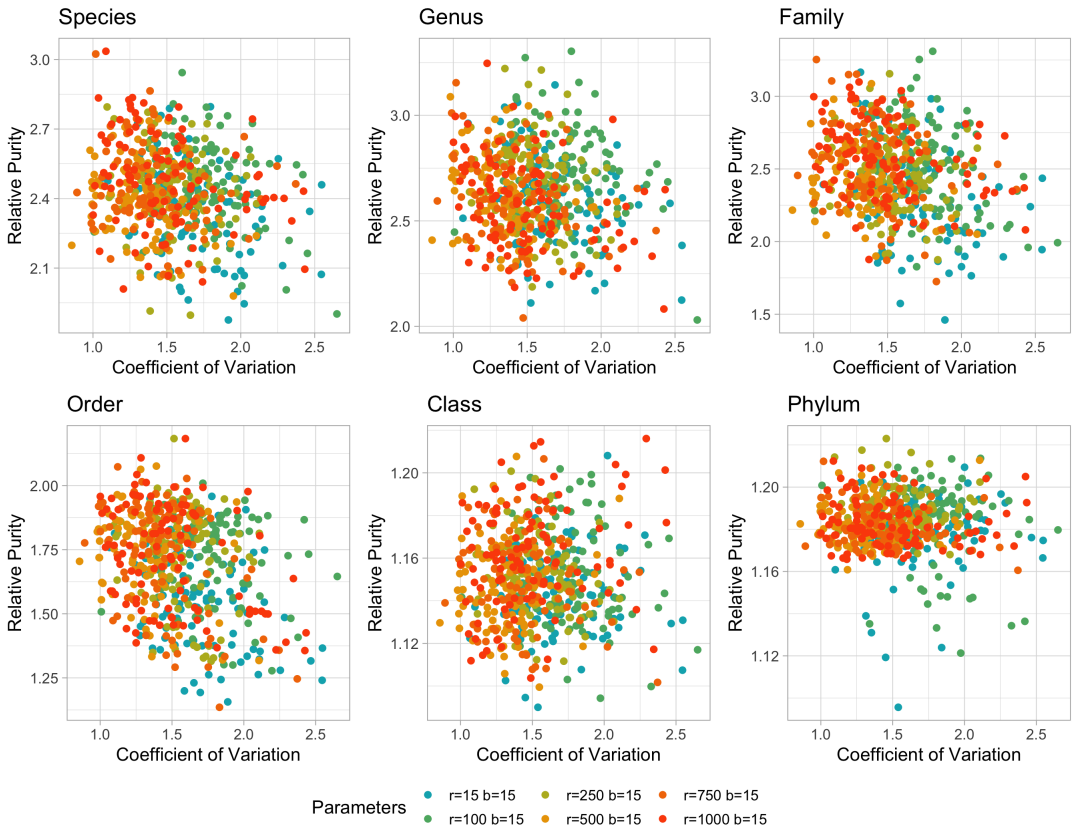

Figure S18. *Relative purity* against Coefficient of Variation for differing values of r ($r = 15$, $r = 100$, $r = 250$, $r = 500$, $r = 750$, $r = 1000$) with $b = 15$.

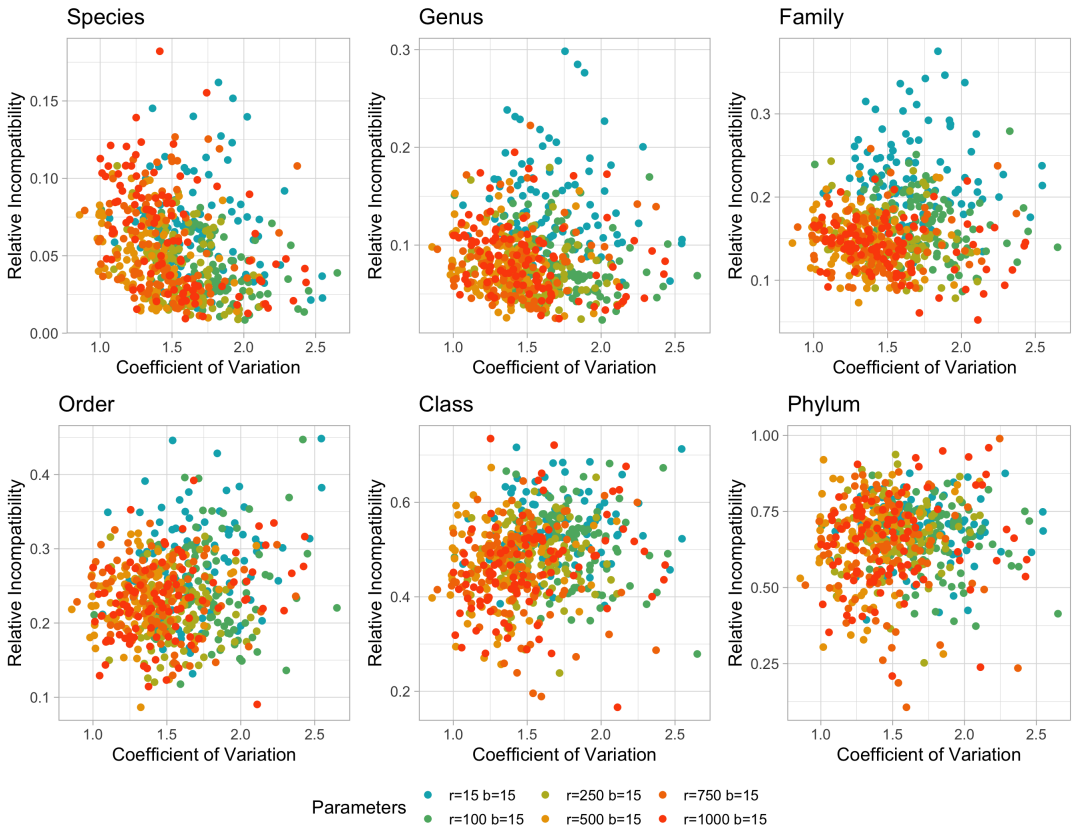

Figure S19. *Relative incompatibility* against Coefficient of Variation for differing values of r ($r = 15$, $r = 100$, $r = 250$, $r = 500$, $r = 750$, and $r = 1000$) with $b = 15$.

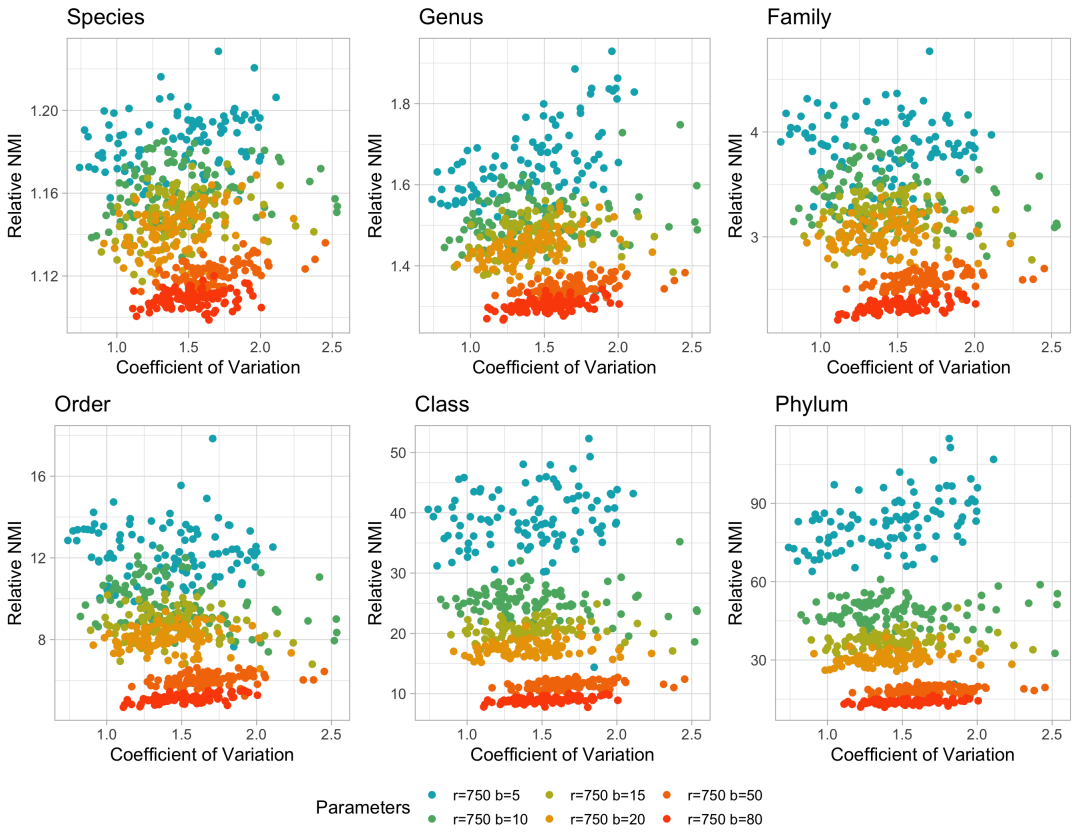

Figure S20. *Relative NMI* against Coefficient of Variation for differing values of b ($b = 5$, $b = 10$, $b = 15$, $b = 20$,$b = 50$, and $b = 80$) with $r = 750$.

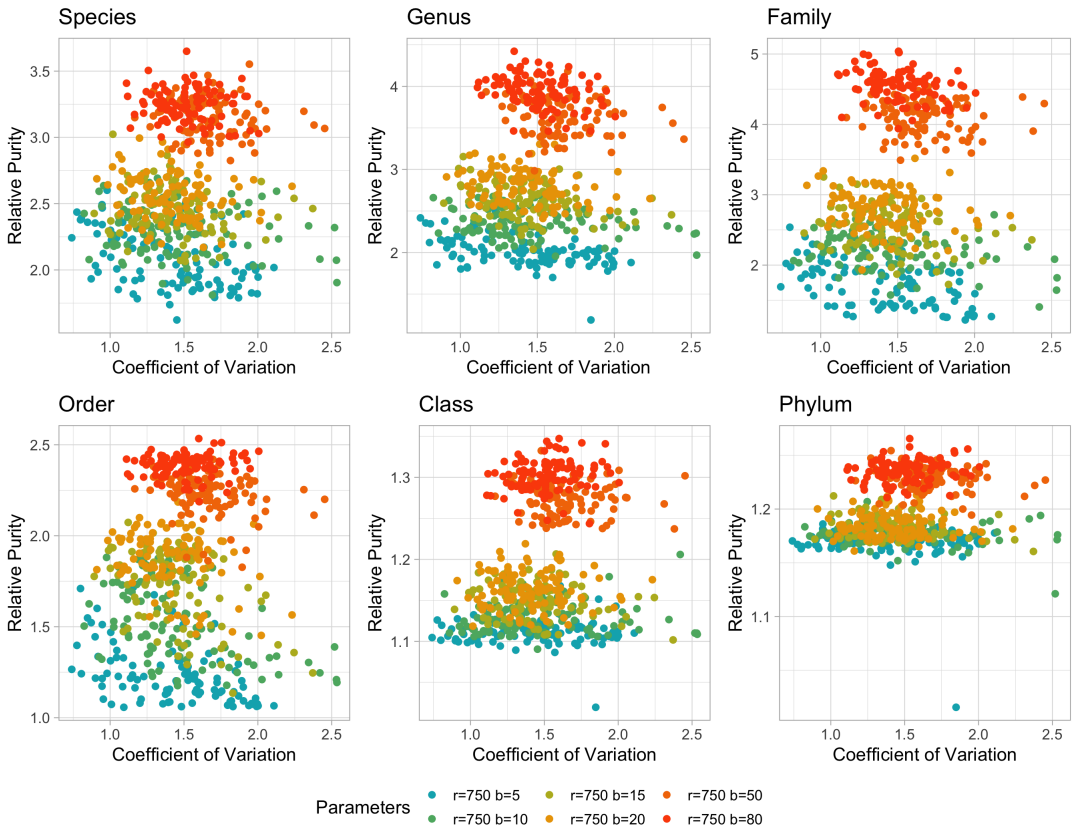

Figure S21. *Relative purity* against Coefficient of Variation for differing values of b ($b = 5$, $b = 10$, $b = 15$, $b = 20$, $b = 50$, and $b = 80$) with $r = 750$.

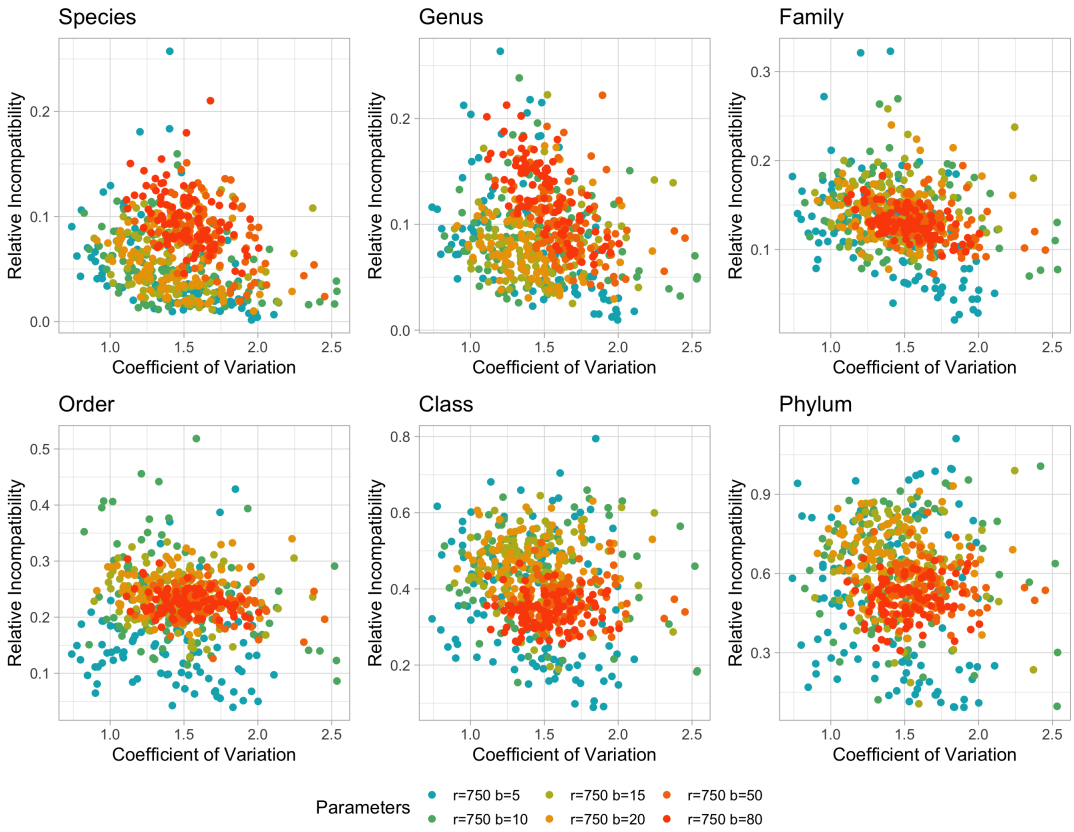

Figure S22. *Relative incompatibility* against Coefficient of Variation for differing values of b ($b = 5$, $b = 10$, $b = 15$, $b = 20$, $b = 50$, and $b = 80$) with $r = 750$.

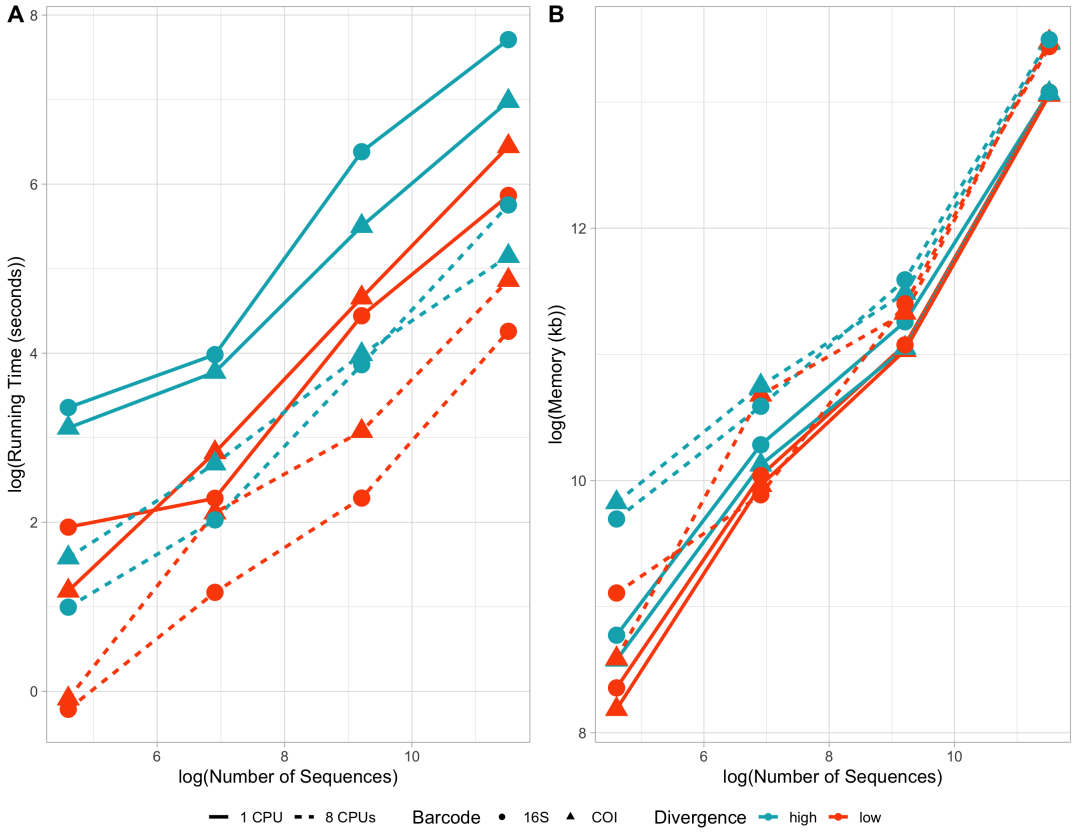

Figure S23. Time (A) and memory (B) requirements of AncestralClust using 100, 1,000, 10,000, and 100,000 sequences of "high" and "low" divergence for 16S and COI.

References

Curd, E. E., Gold, Z., Kandlikar, G. S., Gomer, J., Ogden, M., O'Connell, T., Pipes, L., Schweizer, T. M., Rabichow, L., Lin, M., *et al.* (2019). Anacapa toolkit: an environmental dna toolkit for processing multilocus metabarcode datasets. *Methods in Ecology and Evolution*, **10**(9), 1469–1475.