S3 Appendix

Confusion Matrices for a Single Run of Every Computational Test

Figure 1 (Tests 1-6). Confusion matrices of the assignment that maximizes the accuracy at each taxonomic level of the mtDNA dataset. Predicted labels are numeric cluster assignments, omitted here for readability.

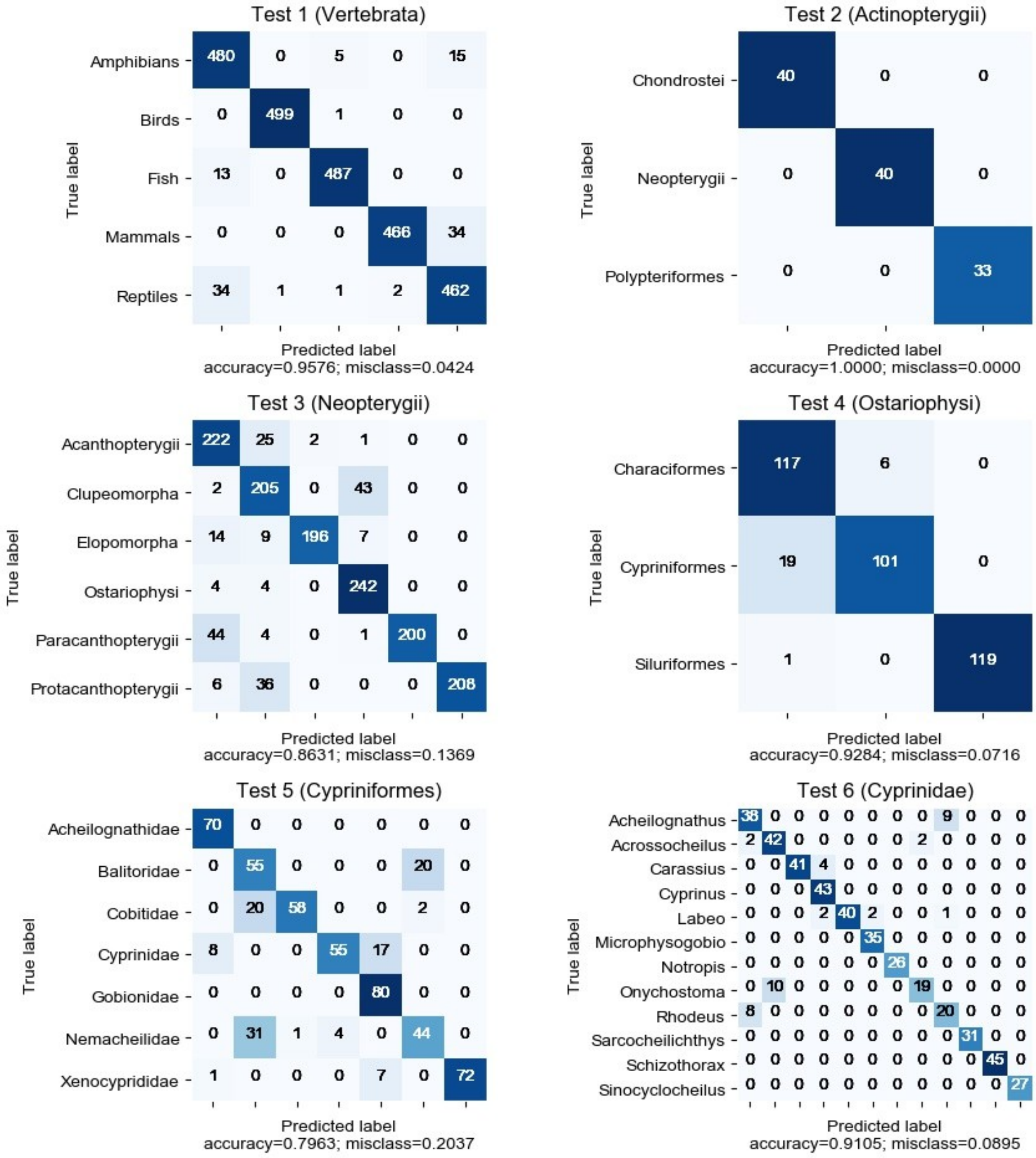


Figure 2 (Tests 7, 8). Confusion matrices of the assignment that maximizes the accuracy for both computational tests with the bacterial dataset at phylum level to families. (Top) All bacterial families

are considered. (Bottom) Only sequences in the phylum Proteobacteria are considered. Predicted labels are numeric cluster assignments, omitted here for readability.

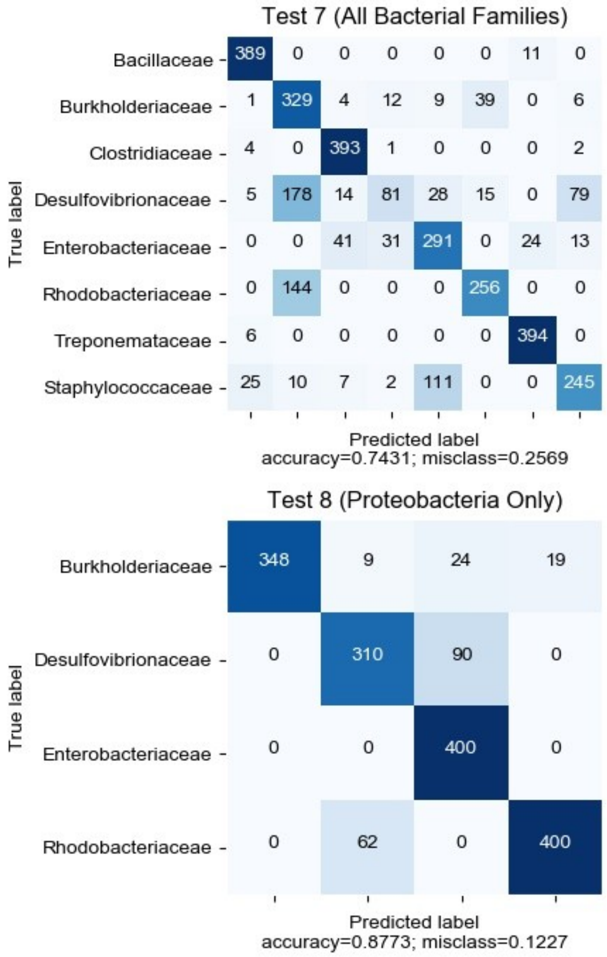
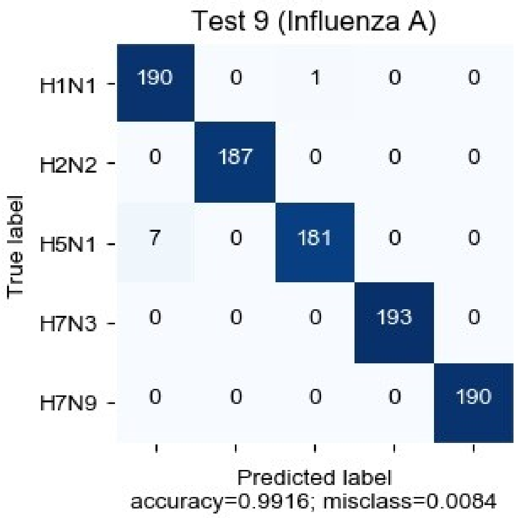
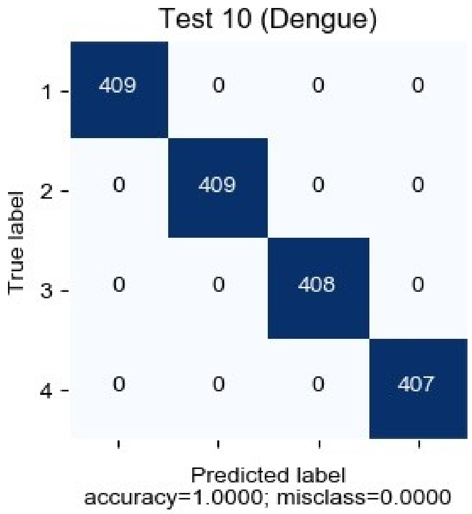
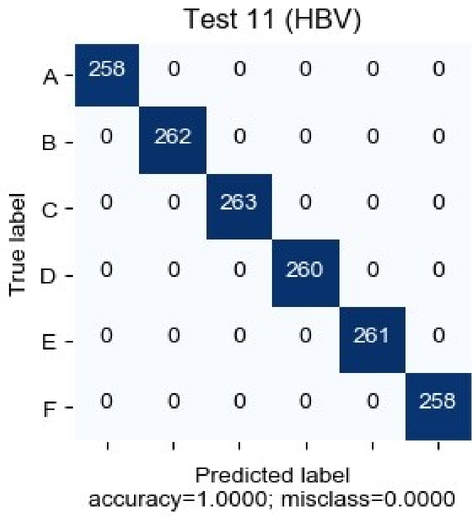


Figure (Tests 9-11). Confusion matrices of the assignment that maximizes the accuracy for the NA-encoding gene of the Influenza A virus, Dengue virus genomes, and HBV genomes. Predicted labels are numeric cluster assignments, omitted here for readability.


Test 9 (Influenza A)
True label
H1N1
H2N2
H5N1
H7N3
H7N9
190 0 1 0 0
0 187 0 0 0
7 0 181 0 0
0 0 0 193 0
0 0 0 0 190
Predicted label
accuracy=0.9916; misclass=0.0084


Test 10 (Dengue)
True label
1
2
3
4
409 0 0 0
0 409 0 0
0 0 408 0
0 0 0 407
Predicted label
accuracy=1.0000; misclass=0.0000


Test 11 (HBV)
True label
A
B
C
D
E
F
258 0 0 0 0 0
0 262 0 0 0 0
0 0 263 0 0 0
0 0 0 260 0 0
0 0 0 0 261 0
0 0 0 0 0 258
Predicted label
accuracy=1.0000; misclass=0.0000