Figure S1 – Simplified overview of the initial filtering of the dataset (A) and machine learning stage (B).
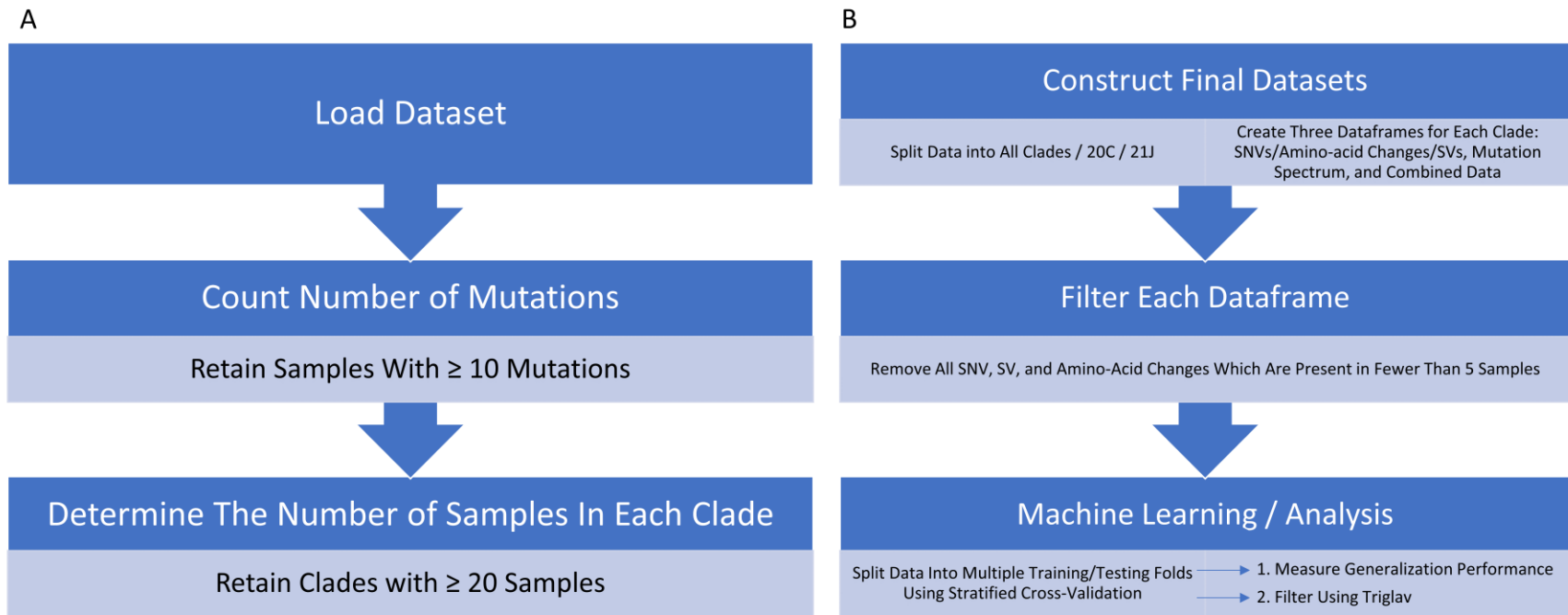
A

Load Dataset

↓

Count Number of Mutations
Retain Samples With ≥ 10 Mutations

↓

Determine The Number of Samples In Each Clade
Retain Clades with ≥ 20 Samples

B

Construct Final Datasets	
Split Data into All Clades / 20C / 21J	Create Three Dataframes for Each Clade: SNVs/Amino-acid Changes/SVs, Mutation Spectrum, and Combined Data

↓

Filter Each Dataframe
Remove All SNV, SV, and Amino-Acid Changes Which Are Present in Fewer Than 5 Samples

↓

Machine Learning / Analysis
Split Data Into Multiple Training/Testing Folds Using Stratified Cross-Validation → 1. Measure Generalization Performance → 2. Filter Using Triglav

Figure S2 – Example of clustering performed by Triglav. The dendrogram visualizes feature clusters for Clade 21J. This dendrogram was produced with Triglav and unsupervised Extremely Randomized Trees. Eight hundred and forty-nine clusters are shown in this dendrogram. On average, each cluster contains 4 features.
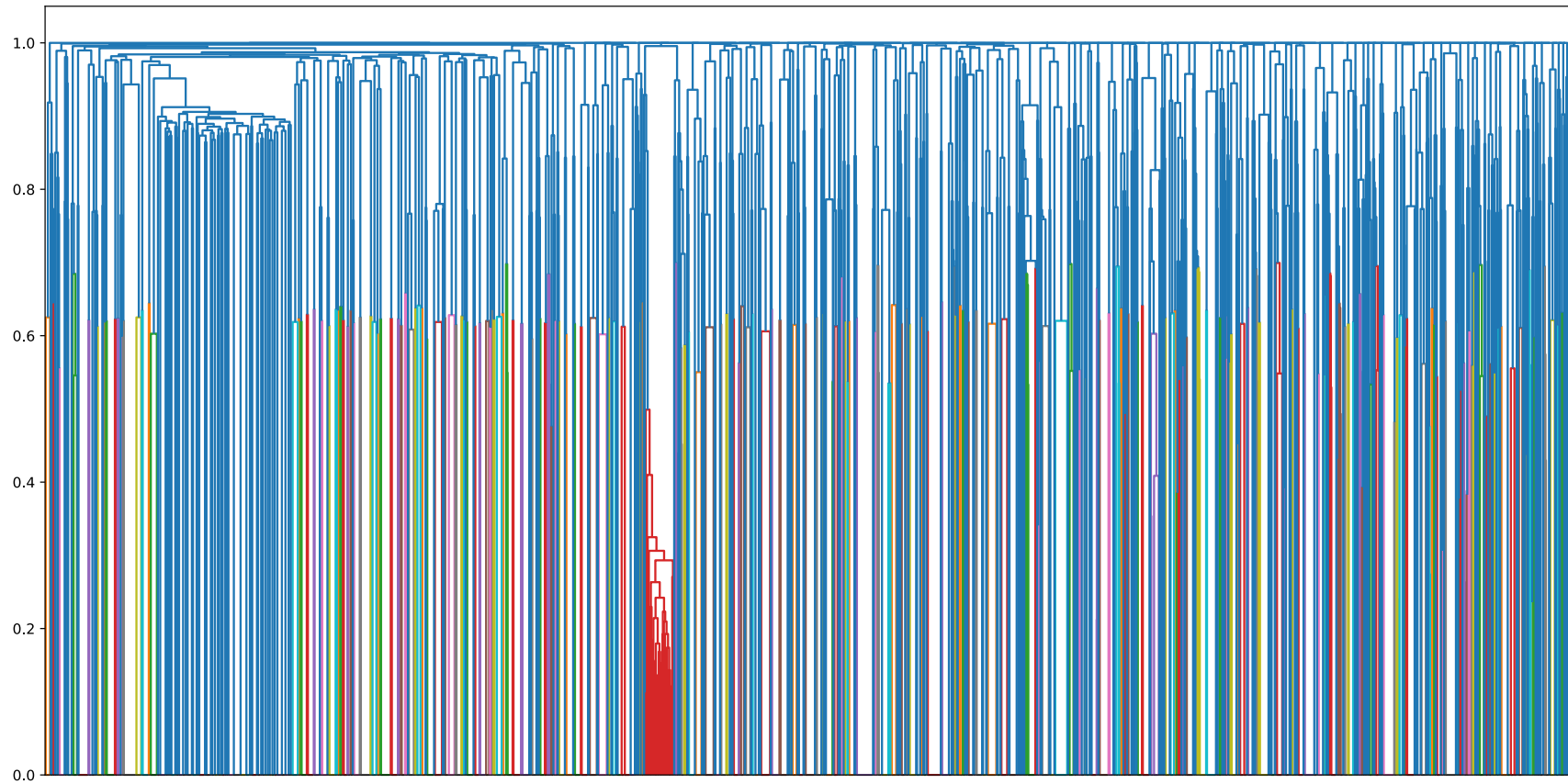
Figure S3 – A bi-clustering of clade 21J. (A) The bi-clustering shows the relationship between deer, human, and mink sequences based on their genome-wide amino-acid changes relative to the reference genome, Wuhan-Hu-1 (MN908947.3). Changes in each gene are colored differently. Rows (samples) and columns (features) were clustered separately using hierarchical clustering and average linkage with distances calculated using the Hamming metric.
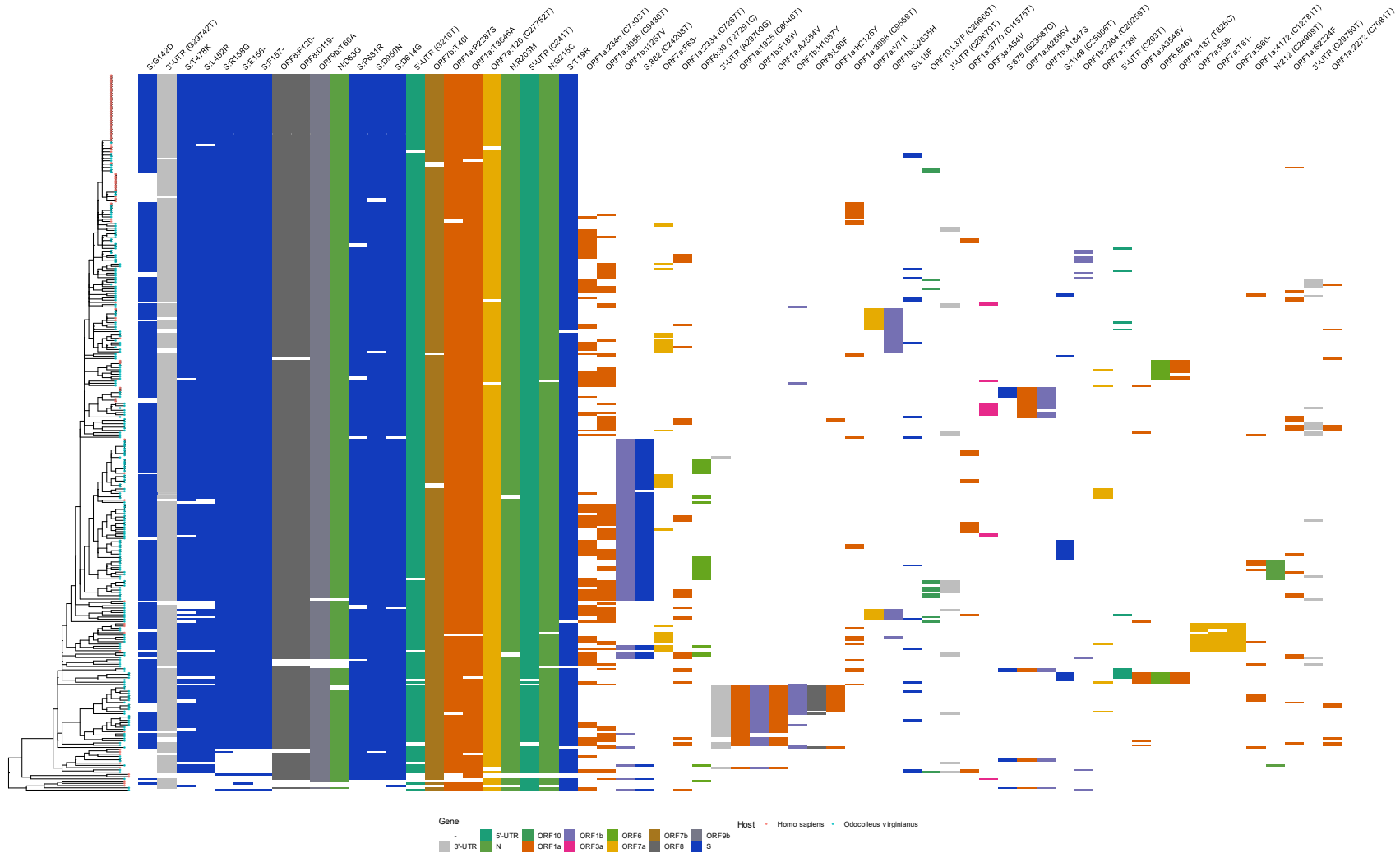
Figure S4 - A bi-clustering of the 'All Clades' dataset using features from the selection analysis not found with Triglav. (A) The bi-clustering shows the relationship between deer, human, and mink sequences based on their genome-wide amino-acid changes relative to the reference genome, Wuhan-Hu-1 (MN908947.3). Changes in each gene are colored differently. Rows (samples) and columns (features) were clustered separately. MAFFT and IQTREE2 were used to create a phylogeny to demonstrate the relationship between samples while hierarchical clustering of features was conducted using average linkage alongside distances calculated using the Hamming metric.