

Figure S: Rank distribution for features over 24 rounds of cross validation (CV). Motifs were discovered for 50% of the available data and in each round of CV features were ranked using information gain. The ranked features are available in Table S1 (Motifs24CV). Ranks for all features for all folds are also available in Table S1 (Ranked Features).

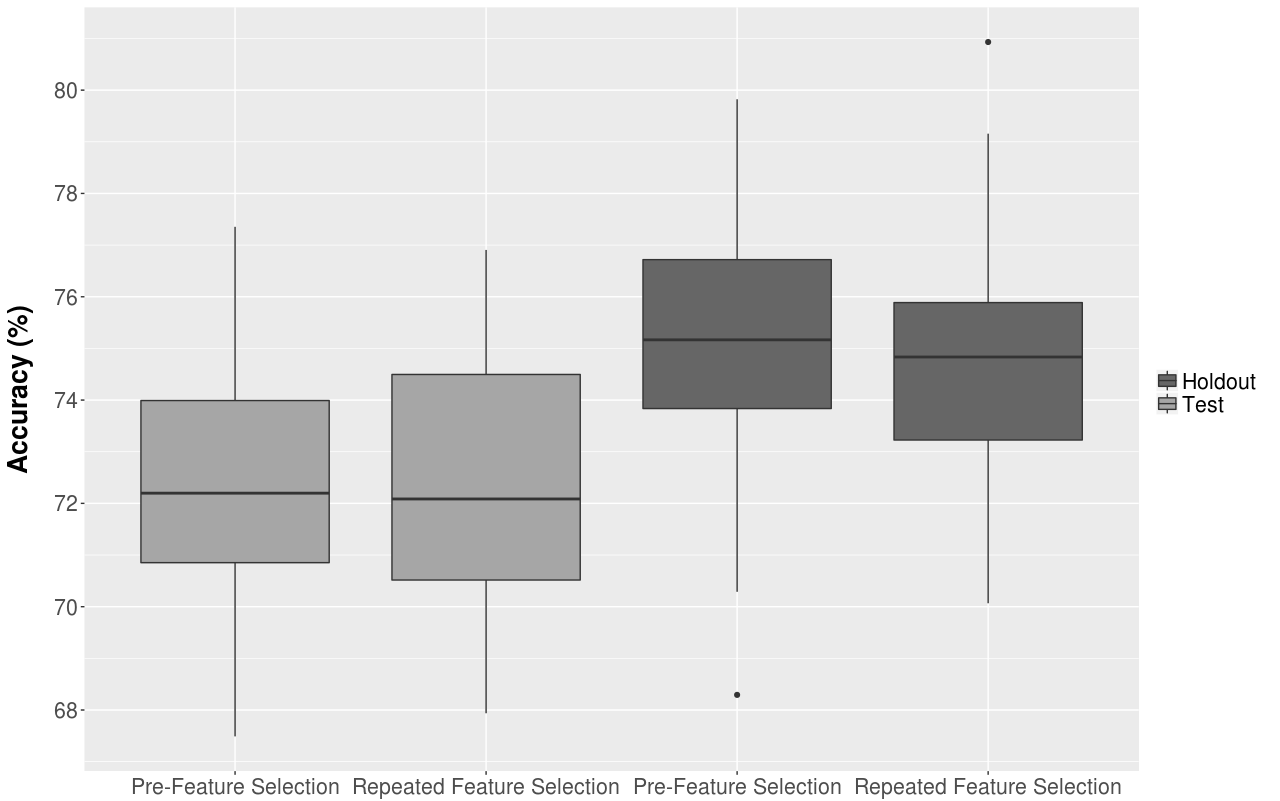


Figure S2: Accuracy distribution for feature selection prior to training and testing and using a strategy where feature selection is repeated for each training-testing iteration.

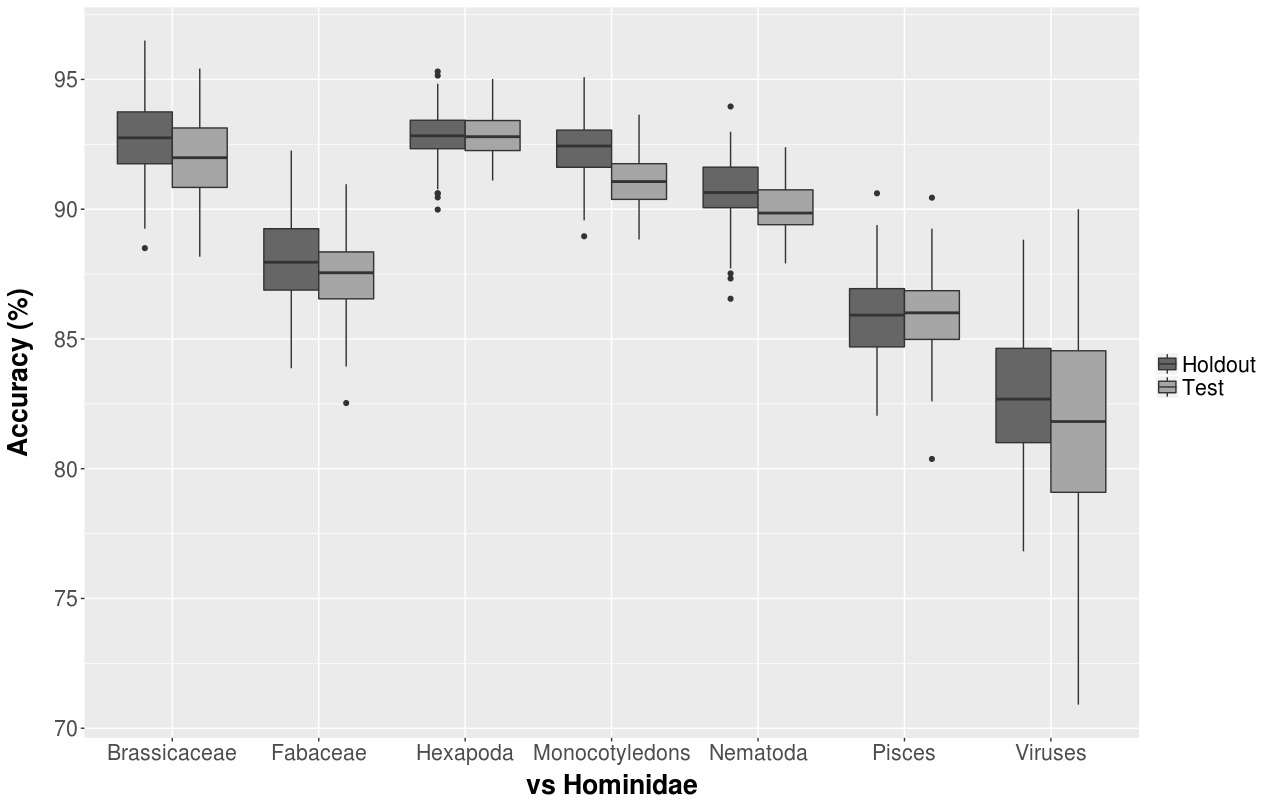


Figure S: Accuracy distribution for 100 fold MCCV for selected species (and groups) against hominidae.

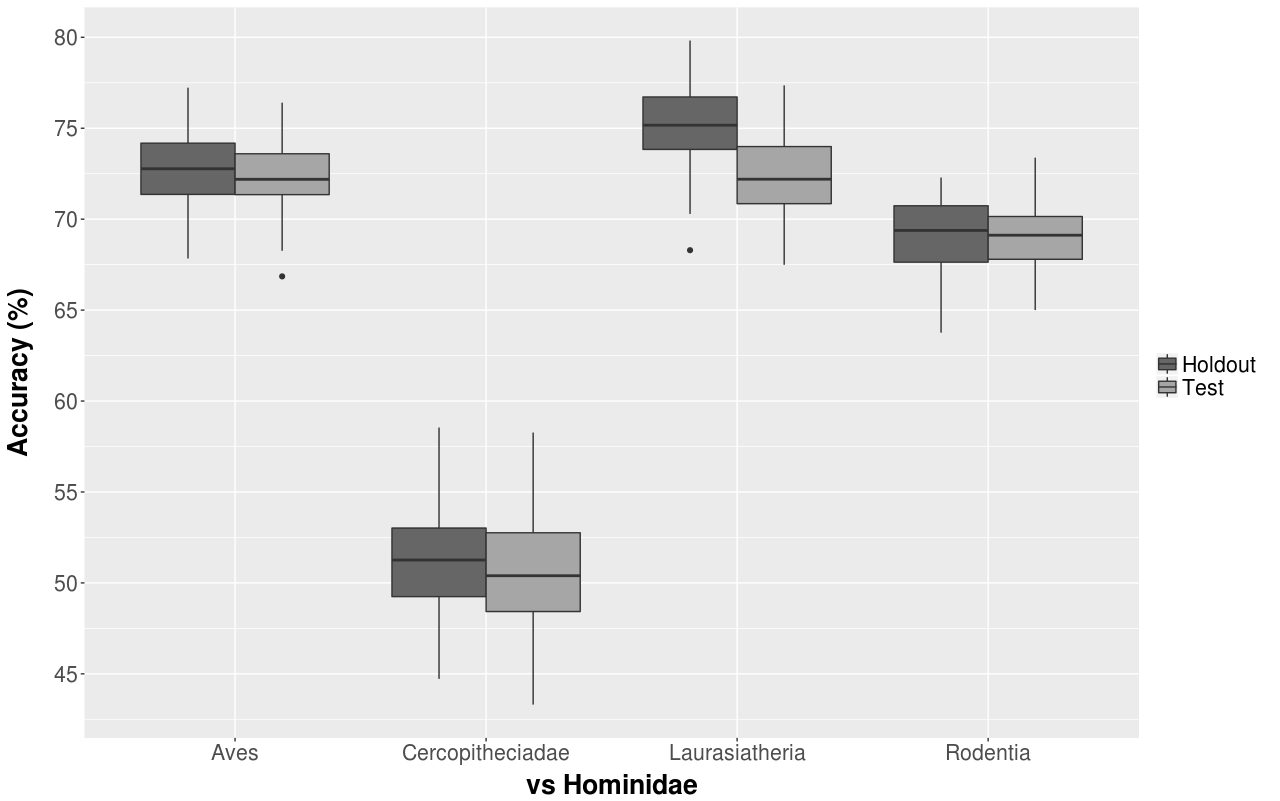


Figure S: Accuracy distribution for 100 fold MCCV for selected species (and groups) against hominidae.

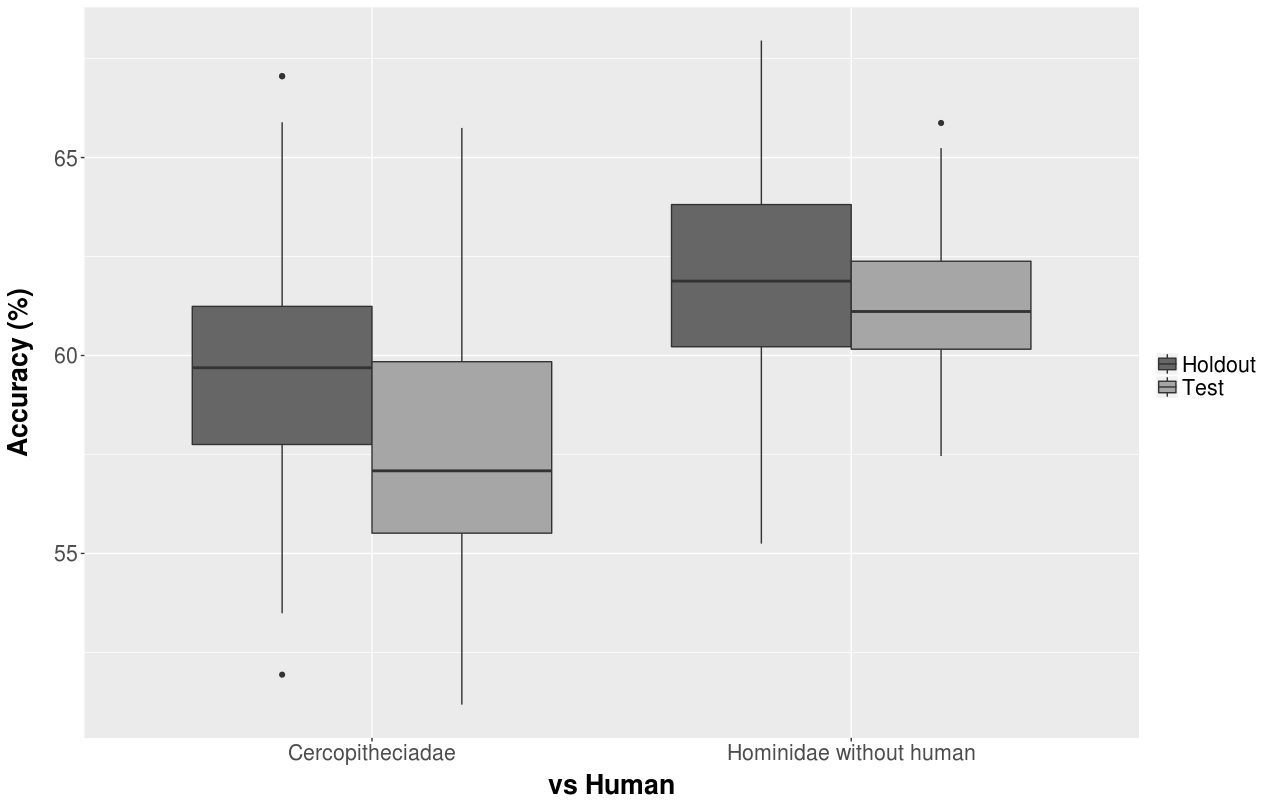


Figure S: Accuracy distribution for 100 fold MCCV for Cercopitheciadae and hominidae against human.

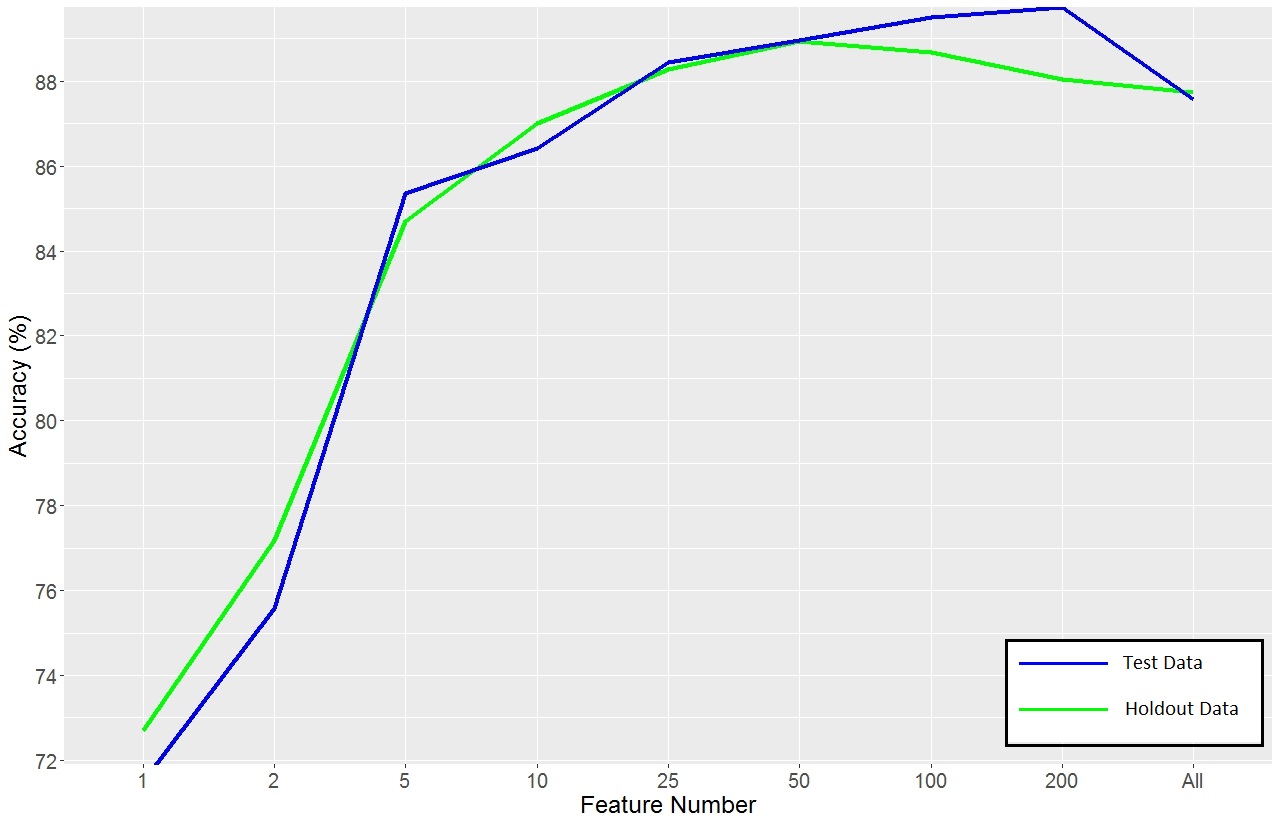


Figure S: Features were sorted by information gain (decreasingly). Feature sets from 1 to all features were created incorporating features with higher information gain first. Then models were created for each feature set and were tested on test and holdout data. The results are presented. For holdout data more than 100 features shouldn’t be used while for test data 100 features almost presents the best selection.