# Summary of Filtered DataFrames and Resulting Matrices

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| DataFrame | Description | 0 Sum Rows Removed | Rows Remaining | Columns Remaining | Notes |
| otu16m\_df | 16S MiSeq | 3,638 | 7,990 | 343 |  |
| otu16n\_df | 16S Nano | 2,474 | 9,154 | 344 |  |
| otu18m\_df | 18S MiSeq | 1,586 | 11,868 | 344 |  |
| otu18n\_df | 18S Nano | 10,292 | 11,868 | 340 |  |
| otu16ren\_df | 16S Rena | 0 | 684 | 23 | No empty rows |
| otu18ren\_df | 18S Rena | \*Unknown\* | 200 | 23 |  |

# Removed host DNA

**16S data**:  
Removed ESVs identified as *chloroplasts* (Order) or *mitochondria* (Family).

**18S data**:  
Removed ESVs classified as *Bacteria*, *Archaea*, or with missing Kingdom.

**18S host DNA cleanup**:  
Further excluded ESVs from *fish* (Actinopterygii), *birds* (Aves), and *mammals* (Mammalia).

# Assessed read length averages before and after host removal

Reads changed drasticallt for 18s , less for 16s

A graph of different colored bars

AI-generated content may be incorrect.

# Assessed and removed low read counts

Tried manual cut offs (20, 50, 500) and an outlier (IQR\_val <- Q3 - Q1, and lower\_bound <- Q1 - 1.5 \* IQR\_val)

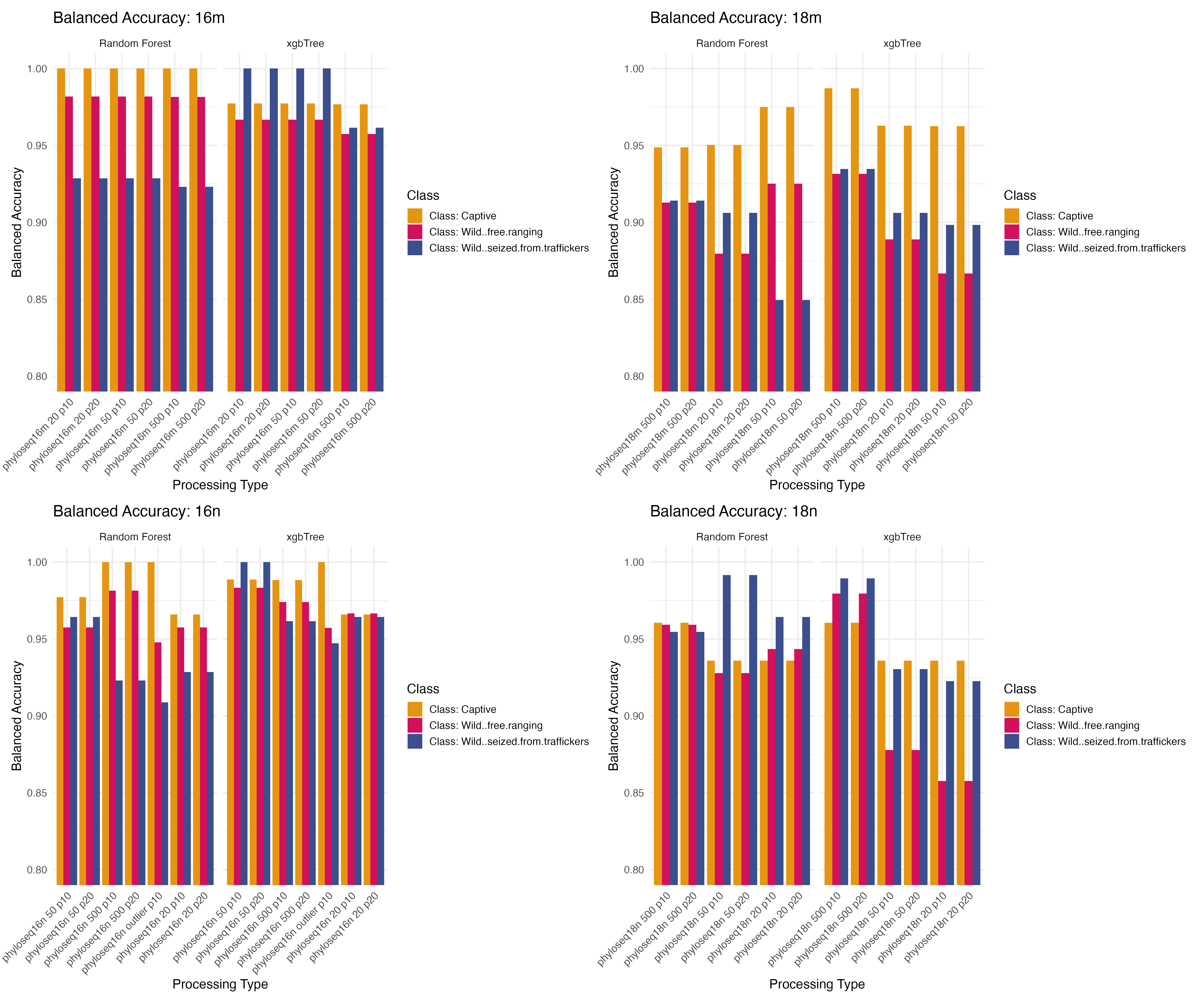
Saved all these dfs. (See coverage plots pdf)

# Assessed and removed low presence cut offs for all dfs made in step 4

Remove ESVs not seen more than 2 times in at least 10% or 20% of the samples.

Save all these dataframes as new phyloseq objects (48 in total)

# Rerun all RF and xgbTree classification models on all the qc DFs and check performance



A screenshot of a test results

AI-generated content may be incorrect.