Machine learning for ecosystem assessment in the Amazon, using environmental DNA data

In Collaboration with NatureMetrics

by

Adamos Spanashis

Supervised by

Dr Ben Calderhead and Dr Oliver Ratmann

September 17, 2019

Contents

Motivation

Why monitor the environment

- Rising human populations.
- Increase in unmanaged waste, especially running waters.
- Assessment of environmental health developed as a response.

Why monitor the environment

- Rising human populations.
- Increase in unmanaged waste, especially running waters.
- Assessment of environmental health developed as a response.

Why monitor the environment

- Rising human populations.
- Increase in unmanaged waste, especially running waters.
- Assessment of environmental health developed as a response.

Traditional environmental monitoring

- Selecting indicator species associated with specific kind of pollution.
- Morpho-taxonomic identification of organisms.
- Filling up biotic indices to quantify pollution.

Traditional environmental monitoring

- Selecting indicator species associated with specific kind of pollution.
- Morpho-taxonomic identification of organisms.
- Filling up biotic indices to quantify pollution.

Traditional environmental monitoring

- Selecting indicator species associated with specific kind of pollution.
- Morpho-taxonomic identification of organisms.
- Filling up biotic indices to quantify pollution.

Traditional monitoring disadvantages

- Taxonomic identification requires experts, is time consuming and expensive.
- Selection of indicator species is arbitrary.
- Taxonomic resolution is low.

Traditional monitoring disadvantages

- Taxonomic identification requires experts, is time consuming and expensive.
- Selection of indicator species is arbitrary.
- Taxonomic resolution is low.

Traditional monitoring disadvantages

- Taxonomic identification requires experts, is time consuming and expensive.
- Selection of indicator species is arbitrary.
- Taxonomic resolution is low.

Traditional monitoring disadvantages

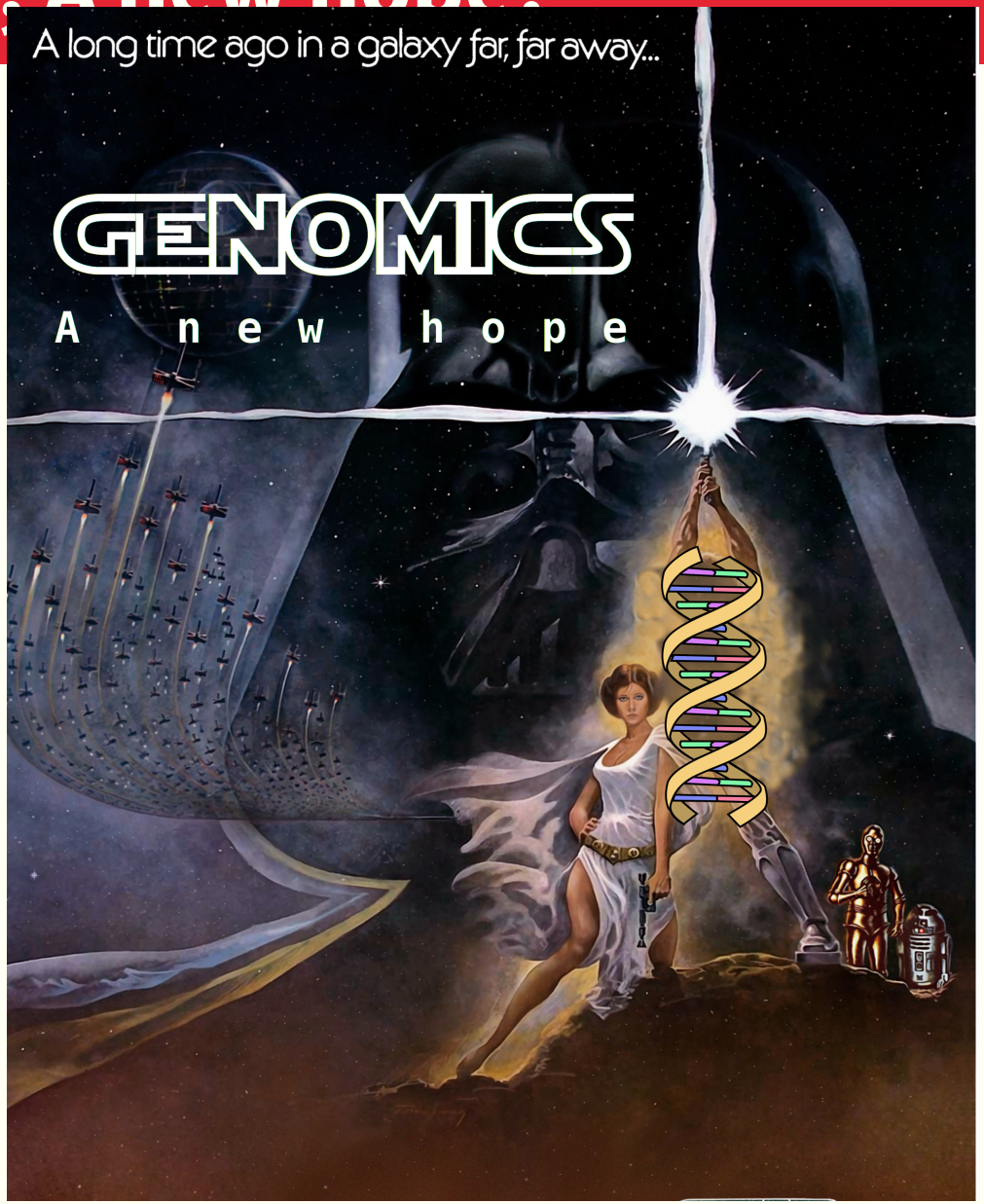
Chironomus Plumosus

Chironomus Zealandicus

Genomics; A new hope?

- Sanger Sequencing 1977
- Barcoding and taxonomic reference libraries
- Next Generation Sequencing
- Costs go down by orders of magnitude

Genomics: A new hope?



Genomics; A new hope?

- Sanger Sequencing 1977
- Barcoding and taxonomic reference libraries
- Next Generation Sequencing
- Costs go down by orders of magnitude

Genomics; A new hope?

- Sanger Sequencing 1977
- Barcoding and taxonomic reference libraries
- Next Generation Sequencing
- Costs go down by orders of magnitude

Genomics; A new hope?

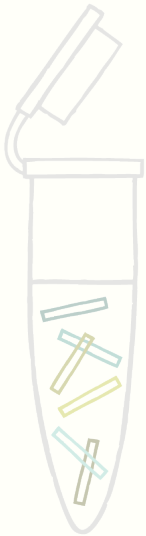
- Sanger Sequencing 1977
- Barcoding and taxonomic reference libraries
- Next Generation Sequencing
- Costs go down by orders of magnitude

Genomics; A new hope?

1

Sampling from the Environment and Isolating the DNA

2



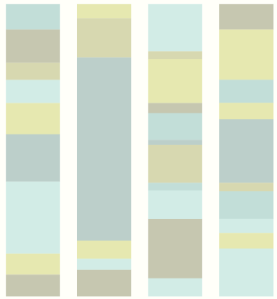
Selecting appropriate primers

3

tttgagtatacaact
ttcgagcatacgact
aacgtccaaaggagt
ttggagcatacgact
aaggtccaaagagt
ttcgagcatacgact
atcgtccaatggagt
aaggtccaaacgagt
aacgtccaaaggagt
tttgagtatacaact

Sequencing DNA metabarcoding libraries

4



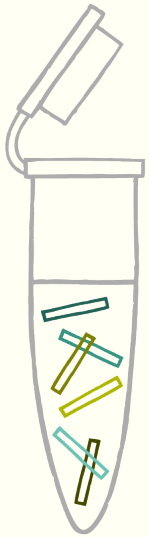
OTU clustering and taxonomic identification

Genomics; A new hope?

1

Sampling from the Environment and Isolating the DNA

2



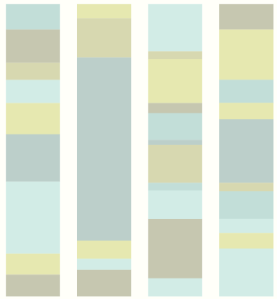
Selecting appropriate primers

3

tttgagtatacaact
ttcgagcatacgact
aacgtccaaaggagt
ttggagcatacgact
aaggtccaaagagt
ttcgagcatacgact
atcgtccaatggagt
aaggtccaaacgagt
aacgtccaaaggagt
tttgagtatacaact

Sequencing DNA metabarcoding libraries

4



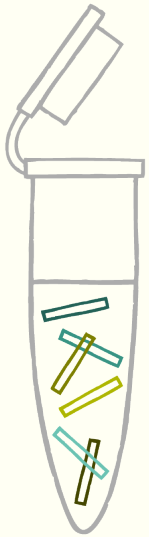
OTU clustering and taxonomic identification

Genomics; A new hope?

1

Sampling from the Environment and Isolating the DNA

2



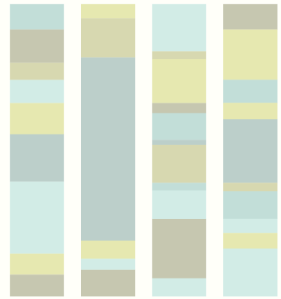
Selecting appropriate primers

3

tttgagtatacaact
ttcgagcatacgact
aacgtccaaaggagt
ttggagcatacgact
aaggtccaaagagt
ttcgagcatacgact
atcgtccaatggagt
aaggtccaaacgagt
aacgtccaaaggagt
tttgagtatacaact

Sequencing DNA metabarcoding libraries

4



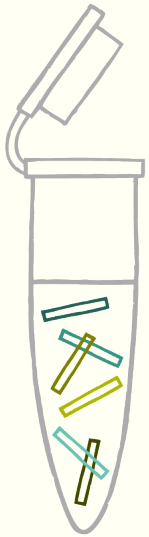
OTU clustering and taxonomic identification

Genomics; A new hope?

1

Sampling from the Environment and Isolating the DNA

2



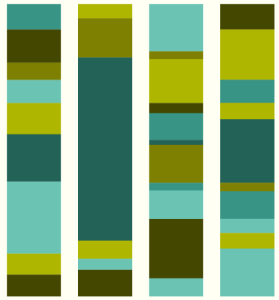
Selecting appropriate primers

3

tttgagtatacaact
ttcgagcatacgact
aacgtccaaaggagt
ttggagcatacgact
aaggtccaaagagt
ttcgagcatacgact
atcgtccaatggagt
aaggtccaaacgagt
aacgtccaaaggagt
tttgagtatacaact

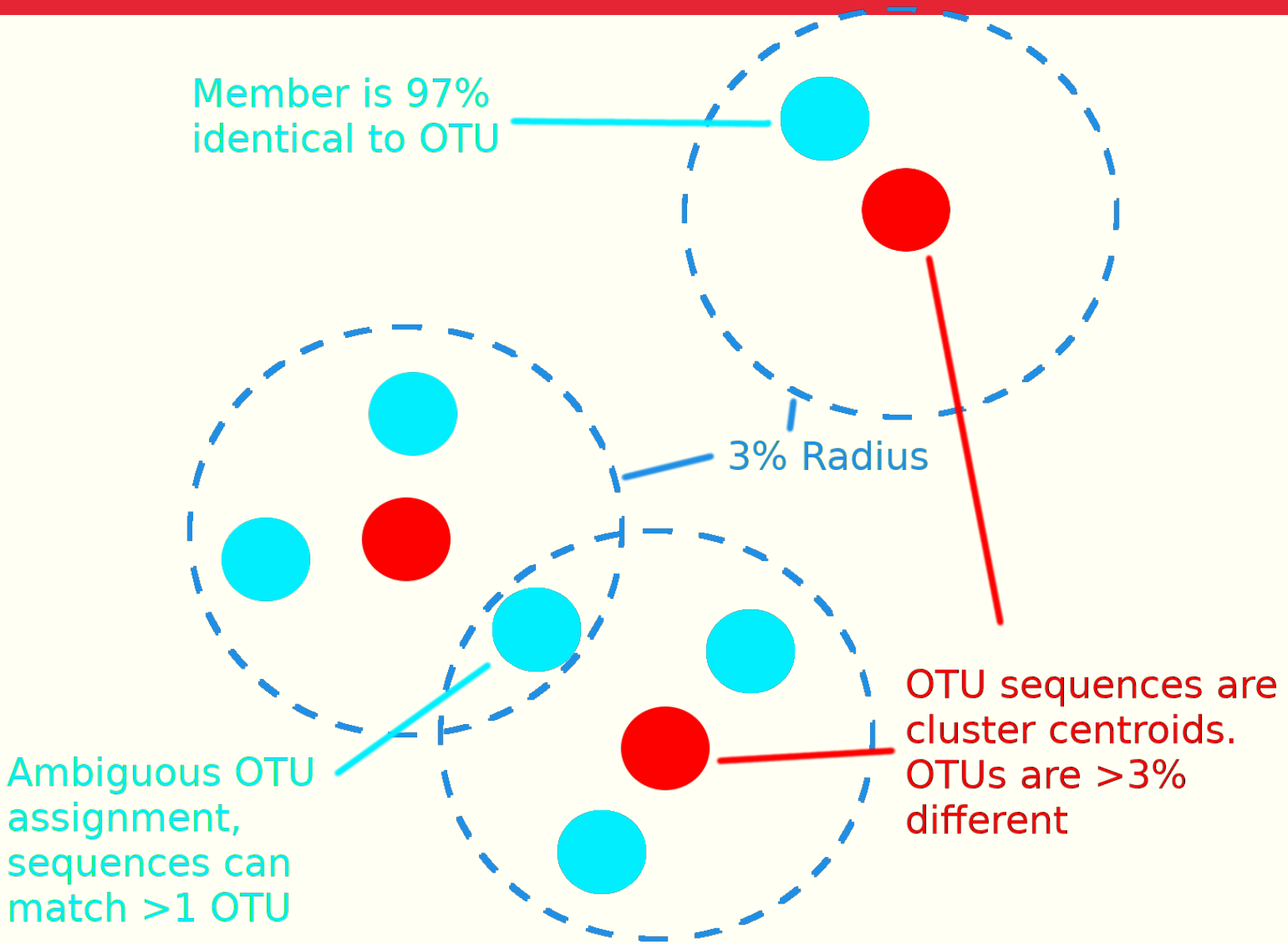
Sequencing DNA metabarcoding libraries

4



OTU clustering and taxonomic identification

Operational Taxonomic Units



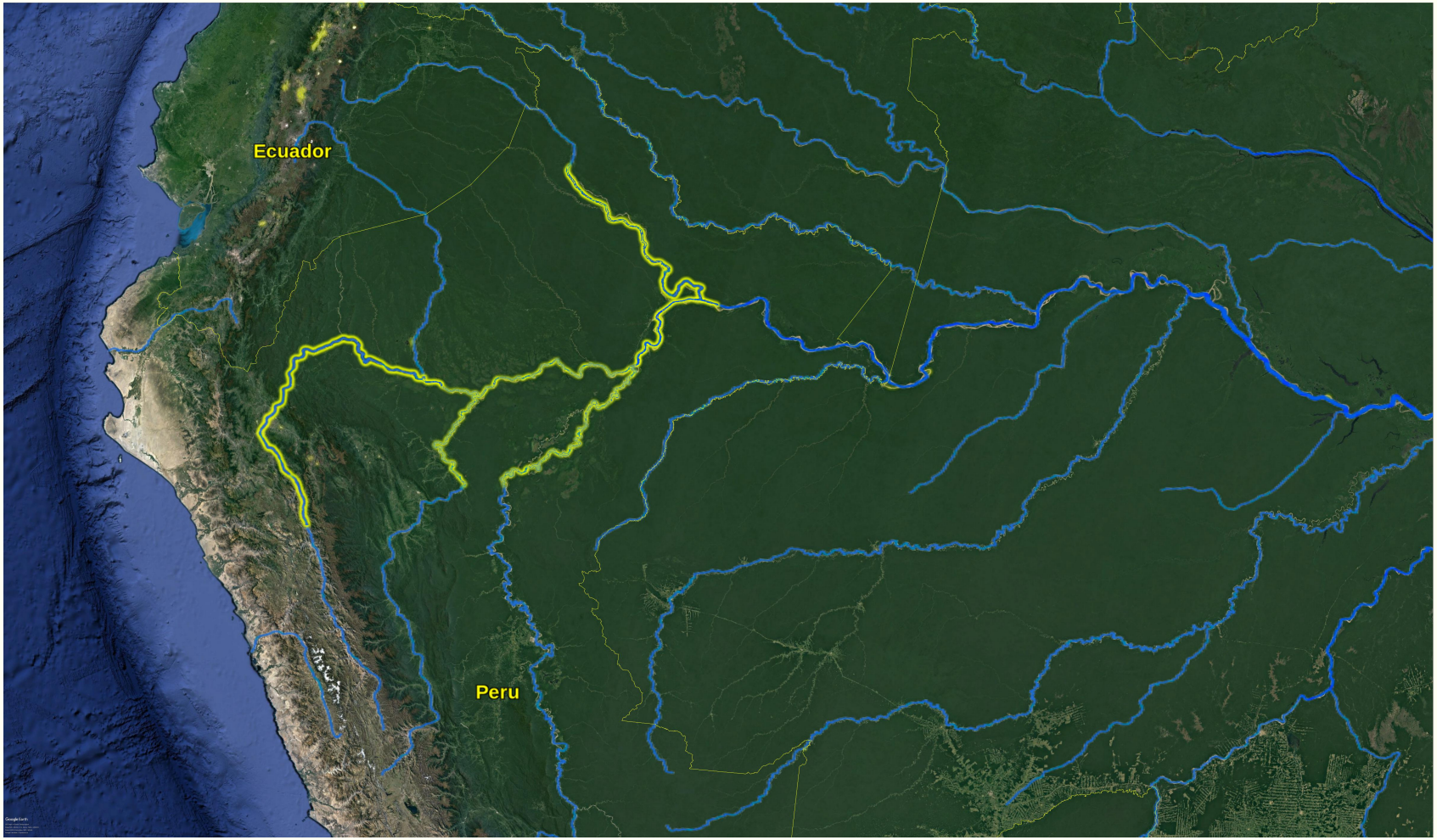
Data

Sampling

- Collected 164 samples from Northern Peruvian rivers
- Sequenced using Next generation sequencing
- 675 OTUs matched with taxonomies from reference libraries
- Discarded the rest
- Recorded colour of water and location

Sampling

Sampling



Sampling

- Collected 164 samples from Northern Peruvian rivers
- Sequenced using Next generation sequencing
- 675 OTUs matched with taxonomies from reference libraries
- Discarded the rest
- Recorded colour of water and location

Sampling

- Collected 164 samples from Northern Peruvian rivers
- Sequenced using Next generation sequencing
- 675 OTUs matched with taxonomies from reference libraries
- Discarded the rest
- Recorded colour of water and location

Sampling

- Collected 164 samples from Northern Peruvian rivers
- Sequenced using Next generation sequencing
- 675 OTUs matched with taxonomies from reference libraries
- Discarded the rest
- Recorded colour of water and location

Sampling

- Collected 164 samples from Northern Peruvian rivers
- Sequenced using Next generation sequencing
- 675 OTUs matched with taxonomies from reference libraries
- Discarded the rest
- Recorded colour of water and location

Meta Data

White vs Black

Meta Data


Peruvian Rivers
9700000
9600000
9500000
9400000
9300000
Northingj
2e+05
4e+05
6e+05
Eastingj
Area_group_name
Huallaga
Maranon_mid
Maranon_lower
Maranon_upper
Ucayali
Tapiche
Napo
Water
Black
White

Features

- Sparse data matrix
- Dimensionality reduction
  - Principle Components Analysis
  - Principle Coordinates Analysis[1]
  - Non-metric Multidimensional Scaling[2]
  - High spearman correlation features removed
- Cumulative sum scaling normalisation and log transformation[3]

[1] Warren S. Torgerson. "Multidimensional scaling: I. Theory and method". In: *Psychometrika* 17.4 (Dec. 1, 1952).

[2] J. B. Kruskal. "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis". In: *Psychometrika* 29.1 (Mar. 1964), pp. 1–27.

[3] Joseph N. Paulson et al. "Robust methods for differential abundance analysis in marker gene surveys". In: *Nature methods* 10.12 (), pp. 1200–1202.

Features

- Sparse data matrix
- Dimensionality reduction
  - Principle Components Analysis
  - Principle Coordinates Analysis[1]
  - Non-metric Multidimensional Scaling[2]
  - High spearman correlation features removed
- Cumulative sum scaling normalisation and log transformation[3]

[1] Warren S. Torgerson. "Multidimensional scaling: I. Theory and method". In: *Psychometrika* 17.4 (Dec. 1, 1952).

[2] J. B. Kruskal. "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis". In: *Psychometrika* 29.1 (Mar. 1964), pp. 1–27.

[3] Joseph N. Paulson et al. "Robust methods for differential abundance analysis in marker gene surveys". In: *Nature methods* 10.12 (), pp. 1200–1202.

Features

- Sparse data matrix
- Dimensionality reduction
  - Principle Components Analysis
  - Principle Coordinates Analysis[1]
  - Non-metric Multidimensional Scaling[2]
  - High spearman correlation features removed
- Cumulative sum scaling normalisation and log transformation[3]

[1] Warren S. Torgerson. "Multidimensional scaling: I. Theory and method". In: *Psychometrika* 17.4 (Dec. 1, 1952).

[2] J. B. Kruskal. "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis". In: *Psychometrika* 29.1 (Mar. 1964), pp. 1–27.

[3] Joseph N. Paulson et al. "Robust methods for differential abundance analysis in marker gene surveys". In: *Nature methods* 10.12 (), pp. 1200–1202.

Features

- Sparse data matrix
- Dimensionality reduction
  - Principle Components Analysis
  - Principle Coordinates Analysis[1]
  - Non-metric Multidimensional Scaling[2]
  - High spearman correlation features removed
- Cumulative sum scaling normalisation and log transformation[3]

[1] Warren S. Torgerson. "Multidimensional scaling: I. Theory and method". In: *Psychometrika* 17.4 (Dec. 1, 1952).

[2] J. B. Kruskal. "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis". In: *Psychometrika* 29.1 (Mar. 1964), pp. 1–27.

[3] Joseph N. Paulson et al. "Robust methods for differential abundance analysis in marker gene surveys". In: *Nature methods* 10.12 (), pp. 1200–1202.

Features

- Sparse data matrix
- Dimensionality reduction
  - Principle Components Analysis
  - Principle Coordinates Analysis[1]
  - Non-metric Multidimensional Scaling[2]
  - High spearman correlation features removed
- Cumulative sum scaling normalisation and log transformation[3]

[1] Warren S. Torgerson. "Multidimensional scaling: I. Theory and method". In: *Psychometrika* 17.4 (Dec. 1, 1952).

[2] J. B. Kruskal. "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis". In: *Psychometrika* 29.1 (Mar. 1964), pp. 1–27.

[3] Joseph N. Paulson et al. "Robust methods for differential abundance analysis in marker gene surveys". In: *Nature methods* 10.12 (), pp. 1200–1202.

Features

- Sparse data matrix
- Dimensionality reduction
  - Principle Components Analysis
  - Principle Coordinates Analysis[1]
  - Non-metric Multidimensional Scaling[2]
  - High spearman correlation features removed
- Cumulative sum scaling normalisation and log transformation[3]

[1] Warren S. Torgerson. "Multidimensional scaling: I. Theory and method". In: *Psychometrika* 17.4 (Dec. 1, 1952).

[2] J. B. Kruskal. "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis". In: *Psychometrika* 29.1 (Mar. 1964), pp. 1–27.

[3] Joseph N. Paulson et al. "Robust methods for differential abundance analysis in marker gene surveys". In: *Nature methods* 10.12 (), pp. 1200–1202.

Features

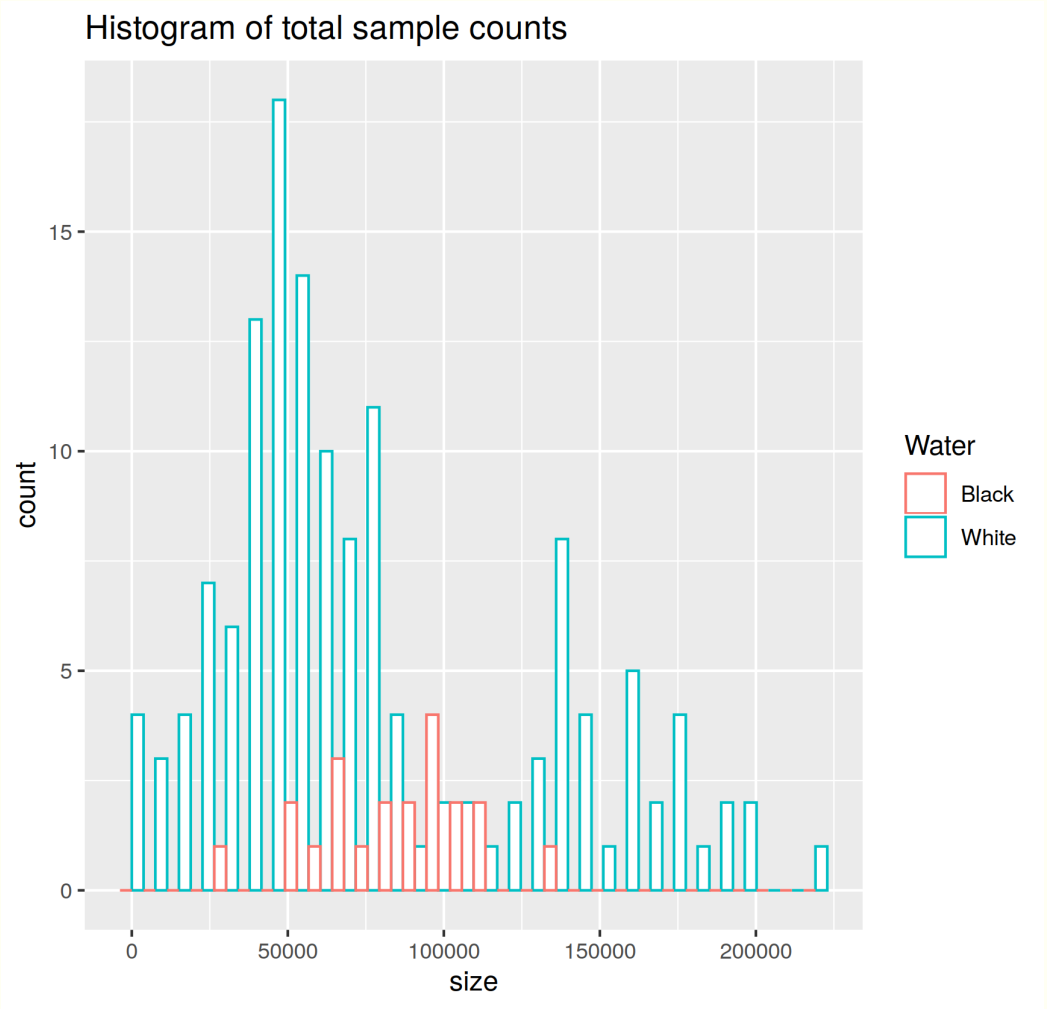
- Sparse data matrix
- Dimensionality reduction
  - Principle Components Analysis
  - Principle Coordinates Analysis[1]
  - Non-metric Multidimensional Scaling[2]
  - High spearman correlation features removed
- Cumulative sum scaling normalisation and log transformation[3]

[1] Warren S. Torgerson. "Multidimensional scaling: I. Theory and method". In: *Psychometrika* 17.4 (Dec. 1, 1952).

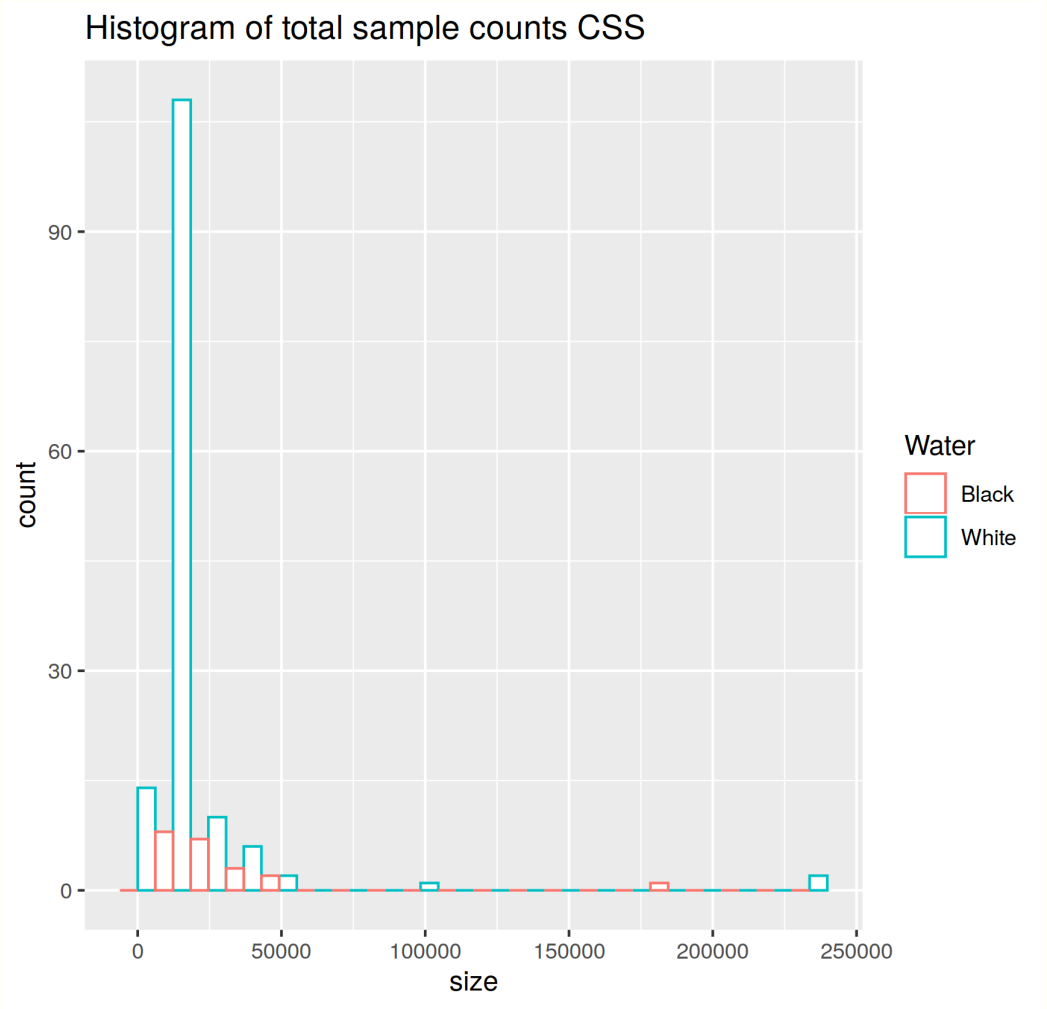
[2] J. B. Kruskal. "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis". In: *Psychometrika* 29.1 (Mar. 1964), pp. 1–27.

[3] Joseph N. Paulson et al. "Robust methods for differential abundance analysis in marker gene surveys". In: *Nature methods* 10.12 (), pp. 1200–1202.

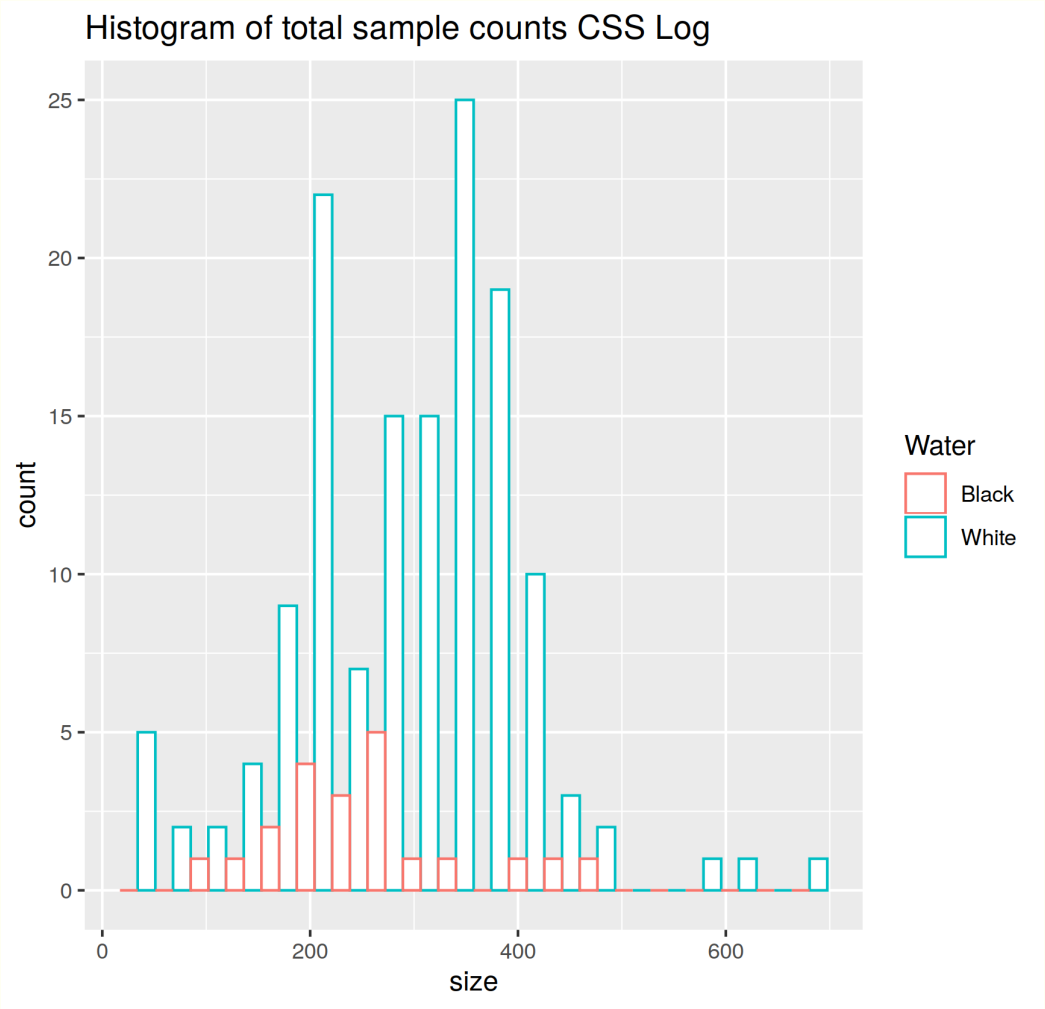
Features


Histogram of total sample counts
count
15
10
5
0
0
50000
100000
150000
200000
size
Water
Black
White

Features


Histogram of total sample counts CSS
count
90
60
30
0
0
50000
100000
150000
200000
250000
size
Water
Black
White

Features


Histogram of total sample counts CSS Log
count
size
0
5
10
15
20
25
0
200
400
600
Water
Black
White

Train-Validation-Test

Location and prediction

- Location of train and test sets will affect performance of classifiers.
- Splitting data simulates different learning conditions .
- Maximally similar
- Maximally dissimilar
- Random with equal distributions of black water in all sets

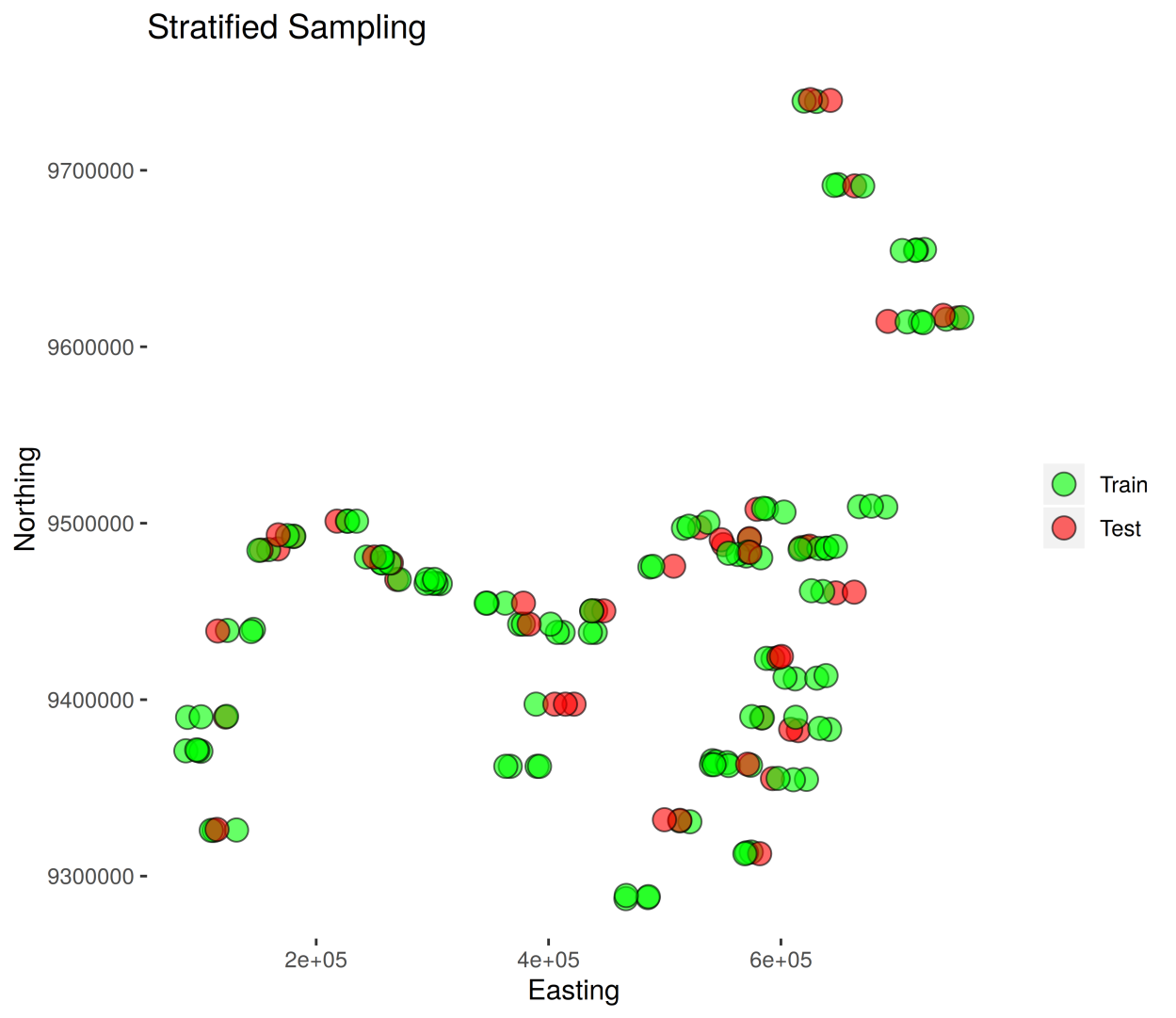
Location and prediction

- Location of train and test sets will affect performance of classifiers.
- Splitting data simulates different learning conditions .
  - Maximally similar
  - Maximally dissimilar
  - Random with equal distributions of black water in all sets

Location and prediction

- Location of train and test sets will affect performance of classifiers.
- Splitting data simulates different learning conditions .
- Maximally similar
- Maximally dissimilar
- Random with equal distributions of black water in all sets

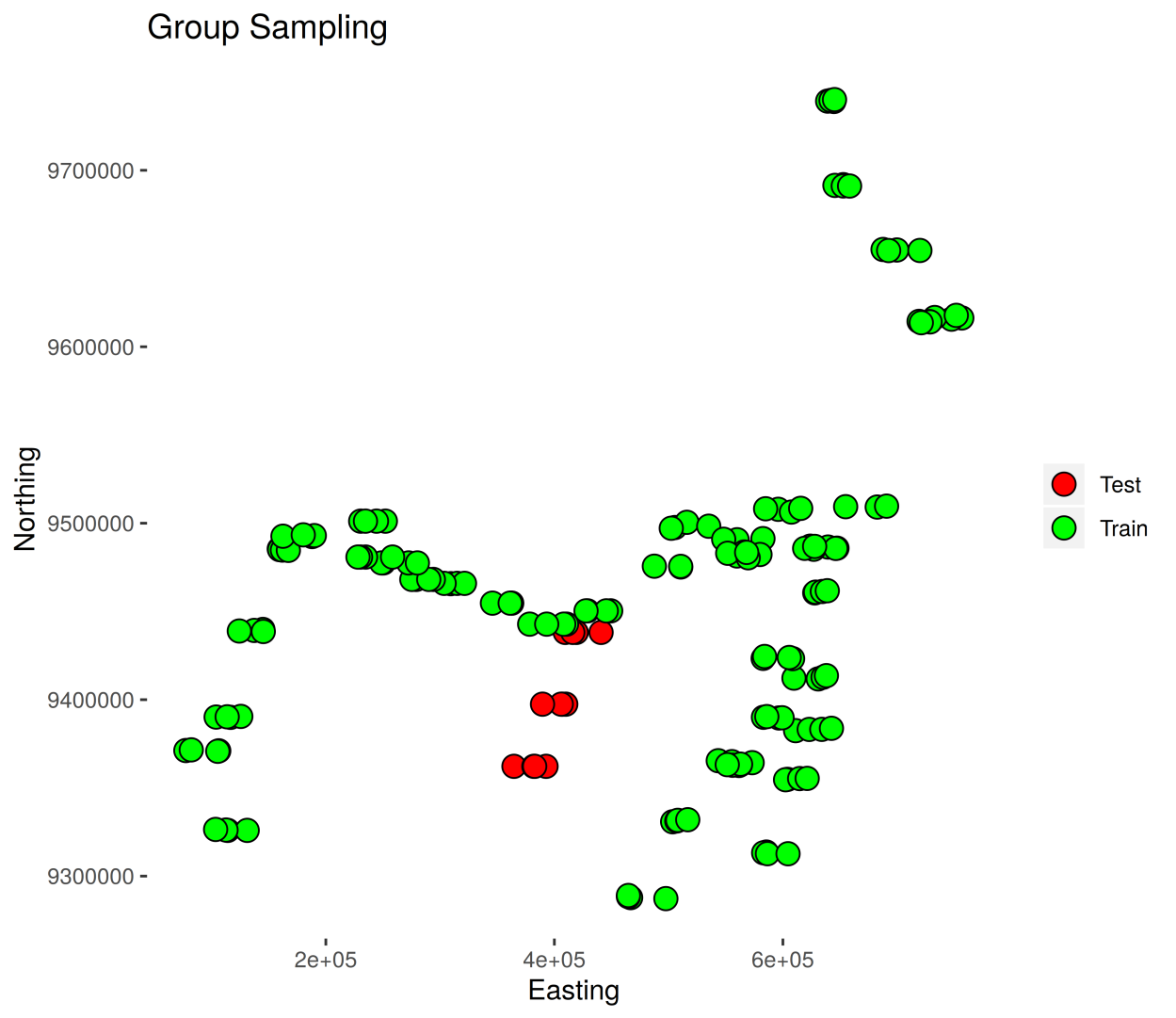
Location and prediction



Location and prediction

- Location of train and test sets will affect performance of classifiers.
- Splitting data simulates different learning conditions .
- Maximally similar
- Maximally dissimilar
- Random with equal distributions of black water in all sets

Location and prediction



Location and prediction

- Location of train and test sets will affect performance of classifiers.
- Splitting data simulates different learning conditions .
- Maximally similar
- Maximally dissimilar
- Random with equal distributions of black water in all sets

Model evaluation

- Split data into train and test sets.
- Cross-validate model on validation sets created from train sets using the same principle as in train-test split; pick best hyperparameters based on accuracy.
- Test on the remaining samples in the test set.
- Repeat until all samples are found in the test set.

Model evaluation

- Split data into train and test sets.
- Cross-validate model on validation sets created from train sets using the same principle as in train-test split; pick best hyperparameters based on accuracy.
- Test on the remaining samples in the test set.
- Repeat until all samples are found in the test set.

Model evaluation

- Split data into train and test sets.
- Cross-validate model on validation sets created from train sets using the same principle as in train-test split; pick best hyperparameters based on accuracy.
- Test on the remaining samples in the test set.
- Repeat until all samples are found in the test set.

Model evaluation

- Split data into train and test sets.
- Cross-validate model on validation sets created from train sets using the same principle as in train-test split; pick best hyperparameters based on accuracy.
- Test on the remaining samples in the test set.
- Repeat until all samples are found in the test set.

Classifiers

- Random Forest
- Logisitic regression with L_1 and L_2 penalty
- Bayesian Logistic Regression with Laplacian and Gaussian priors

Classifiers

- Random Forest
- Logisitic regression with L_1 and L_2 penalty
- Bayesian Logistic Regression with Laplacian and Gaussian priors

Classifiers

- Random Forest
- Logisitic regression with L_1 and L_2 penalty
- Bayesian Logistic Regression with Laplacian and Gaussian priors

Results

Evaluating performance

- Can not use F score or ROC based on black water samples
- Some validation sets do not contain any black waters so most measures are undefined. Accuracy picks better models than F score based on white water.

Results presented in confusion matrix and accuracy:

$$\begin{bmatrix} \text{Black samples Predicted correctly} & \text{Black samples Predicted falsely} \\ \text{White samples Predicted falsely} & \text{White samples Predicted correctly} \end{bmatrix} \tag{1}$$

Evaluating performance

- Can not use F score or ROC based on black water samples
- Some validation sets do not contain any black waters so most measures are undefined. Accuracy picks better models than F score based on white water.

Results presented in confusion matrix and accuracy:

$$\begin{bmatrix} \text{Black samples Predicted correctly} & \text{Black samples Predicted falsely} \\ \text{White samples Predicted falsely} & \text{White samples Predicted correctly} \end{bmatrix} \tag{1}$$

Evaluating performance

- Can not use F score or ROC based on black water samples
- Some validation sets do not contain any black waters so most measures are undefined. Accuracy picks better models than F score based on white water.

Results presented in confusion matrix and accuracy:

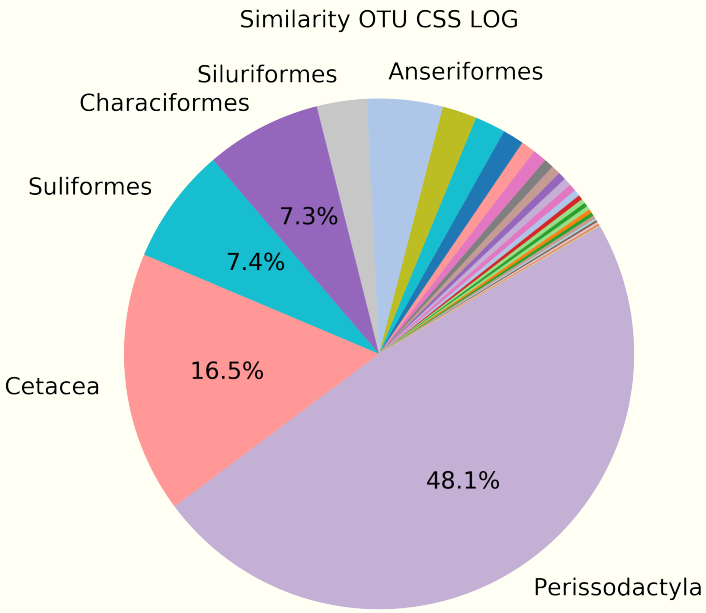
$$\begin{bmatrix} \text{Black samples Predicted correctly} & \text{Black samples Predicted falsely} \\ \text{White samples Predicted falsely} & \text{White samples Predicted correctly} \end{bmatrix} \tag{1}$$

Best of Maximally similar

Features used	Confusion Matrix Predicted Black	Confusion Matrix Predicted White	Accuracy
LR L1 OTU	19 1	2 142	98.17%
LR L1 OTU LOW	19 0	2 143	98.78%
LR L1 OTU CSS	18 0	3 143	98.17%
LR L1 OTU CSS LOG	19 1	2 142	98.17%
LR L2 OTU CSS LOG	20 1	1 142	98.78%
RF OTU LOW	18 3	3 140	96.34%
RF OTU CSS	18 3	3 140	96.34%
RF OTU CSS LOG	19 3	2 140	96.95%

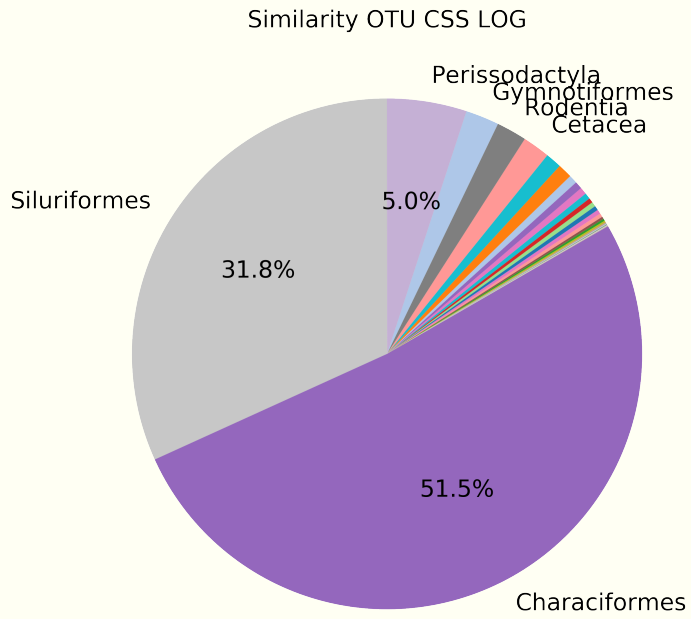
Most important taxonomic orders

CSS LOG.bb



Aggregating importance by averaging over Order

CSS LOG.bb



Aggregating importance by summing over Order

Most important taxonomic orders

South American Tapir[1]

Amazon river dolphin[2]

[1] Sharp Charles. *Tapir*. In: *Wikipedia*. Page Version ID: 913557340. Sept. 1, 2019. URL: https://en.wikipedia.org/w/index.php?title=Tapir&oldid=913557340 (visited on 09/13/2019)

[2] *5 amazing facts about the Amazon pink river dolphin*. Aqua Expeditions. Apr. 13, 2019. URL: https://www.aquaexpeditions.com/blog/amazon/facts-amazon-pink-river-dolphin/ (visited on 09/13/2019)

Best of Maximally Dissimilar

Features used	Confusion Matrix Predicted Black	Confusion Matrix Predicted White	Accuracy
L1 PCoA CSS 99%	8 5	13 138	89.02%
L1 PCoA 90%	7 5	14 138	88.41%
L2 OTU CSS LOG	12 1	9 142	93.90%
RF OTU CSS LOG	7 2	14 141	90.24%
RF OTU CSS	7 2	14 141	90.24%

Bayesian Logistic Regression

- Sampling is very slow and poor when using any OTU or PCoA features.
- Only works with 20 dimensions of NMDS
- Produces better results than Maximum Likelihood version of Logistic Regression using same feature set

Features used	Confusion Matrix		Accuracy
	Predicted Black	Predicted White	
Max. Sim. NMDS L_1	13	8	95.12%
	0	143	
Max. Sim. NMDS L_2	8	13	92.07%
	0	143	
Max. Dissim. NMDS L_1	3	18	83.54%
	9	134	
Max. Dissim. NMDS L_2	0	21	92.07%
	21	140	

Bayesian Logistic Regression

- Sampling is very slow and poor when using any OTU or PCoA features.
- Only works with 20 dimensions of NMDS
- Produces better results than Maximum Likelihood version of Logistic Regression using same feature set

	Confusion Matrix		Accuracy
Features used	Predicted Black	Predicted White	
Max. Sim. NMDS L_1	13 0	8 143	95.12%
Max. Sim. NMDS L_2	8 0	13 143	92.07%
Max. Dissim. NMDS L_1	3 9	18 134	83.54%
Max. Dissim. NMDS L_2	0 21	21 140	92.07%

Bayesian Logistic Regression

- Sampling is very slow and poor when using any OTU or PCoA features.
- Only works with 20 dimensions of NMDS
- Produces better results than Maximum Likelihood version of Logistic Regression using same feature set

	Confusion Matrix		Accuracy
Features used	Predicted Black	Predicted White	
Max. Sim. NMDS L_1	13 0	8 143	95.12%
Max. Sim. NMDS L_2	8 0	13 143	92.07%
Max. Dissim. NMDS L_1	3 9	18 134	83.54%
Max. Dissim. NMDS L_2	0 21	21 140	92.07%

Bayesian Logistic Regression

- Sampling is very slow and poor when using any OTU or PCoA features.
- Only works with 20 dimensions of NMDS
- Produces better results than Maximum Likelihood version of Logistic Regression using same feature set

Features used	Confusion Matrix		Accuracy
	Predicted Black	Predicted White	
Max. Sim. NMDS L_1	13	8	95.12%
	0	143	
Max. Sim. NMDS L_2	8	13	92.07%
	0	143	
Max. Dissim. NMDS L_1	3	18	83.54%
	9	134	
Max. Dissim. NMDS L_2	0	21	92.07%
	21	140	

Conclusions

- Maximum Dissimilarity is a harder setting than similarity for prediction
- CSS normalisation and Log transformation produce the best feature set. Ordination methods may not be useful.
- Only a small set of taxonomic Orders are contributing to prediction
- Between Random Forest and Logistic regression there is no significant difference in predictive ability.
- Bayesian methods are promising but slow

Conclusions

- Maximum Dissimilarity is a harder setting than similarity for prediction
- CSS normalisation and Log transformation produce the best feature set. Ordination methods may not be useful.
- Only a small set of taxonomic Orders are contributing to prediction
- Between Random Forest and Logistic regression there is no significant difference in predictive ability.
- Bayesian methods are promising but slow

Conclusions

- Maximum Dissimilarity is a harder setting than similarity for prediction
- CSS normalisation and Log transformation produce the best feature set. Ordination methods may not be useful.
- Only a small set of taxonomic Orders are contributing to prediction
- Between Random Forest and Logistic regression there is no significant difference in predictive ability.
- Bayesian methods are promising but slow

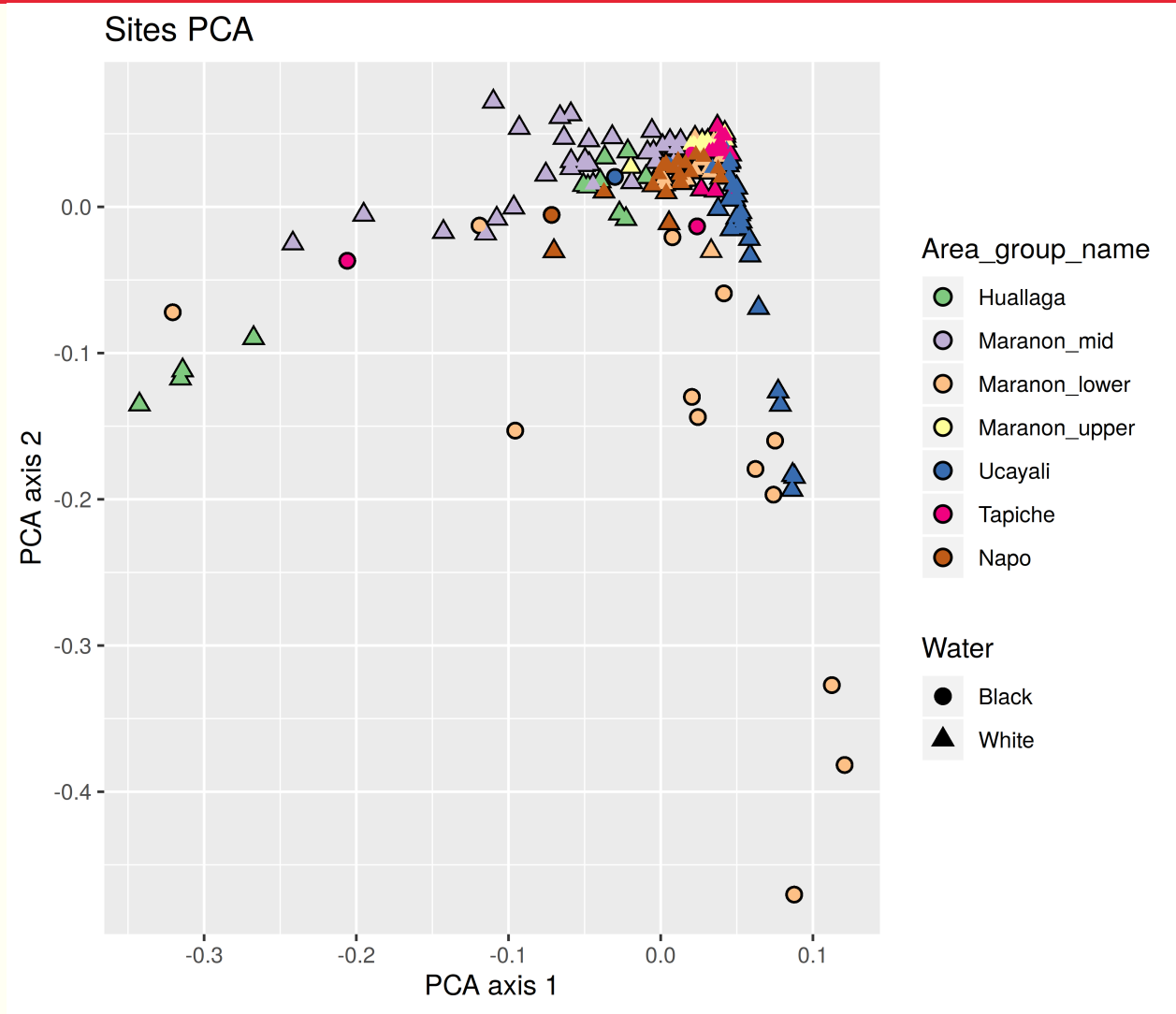
Conclusions

- Maximum Dissimilarity is a harder setting than similarity for prediction
- CSS normalisation and Log transformation produce the best feature set. Ordination methods may not be useful.
- Only a small set of taxonomic Orders are contributing to prediction
- Between Random Forest and Logistic regression there is no significant difference in predictive ability.
- Bayesian methods are promising but slow

Conclusions

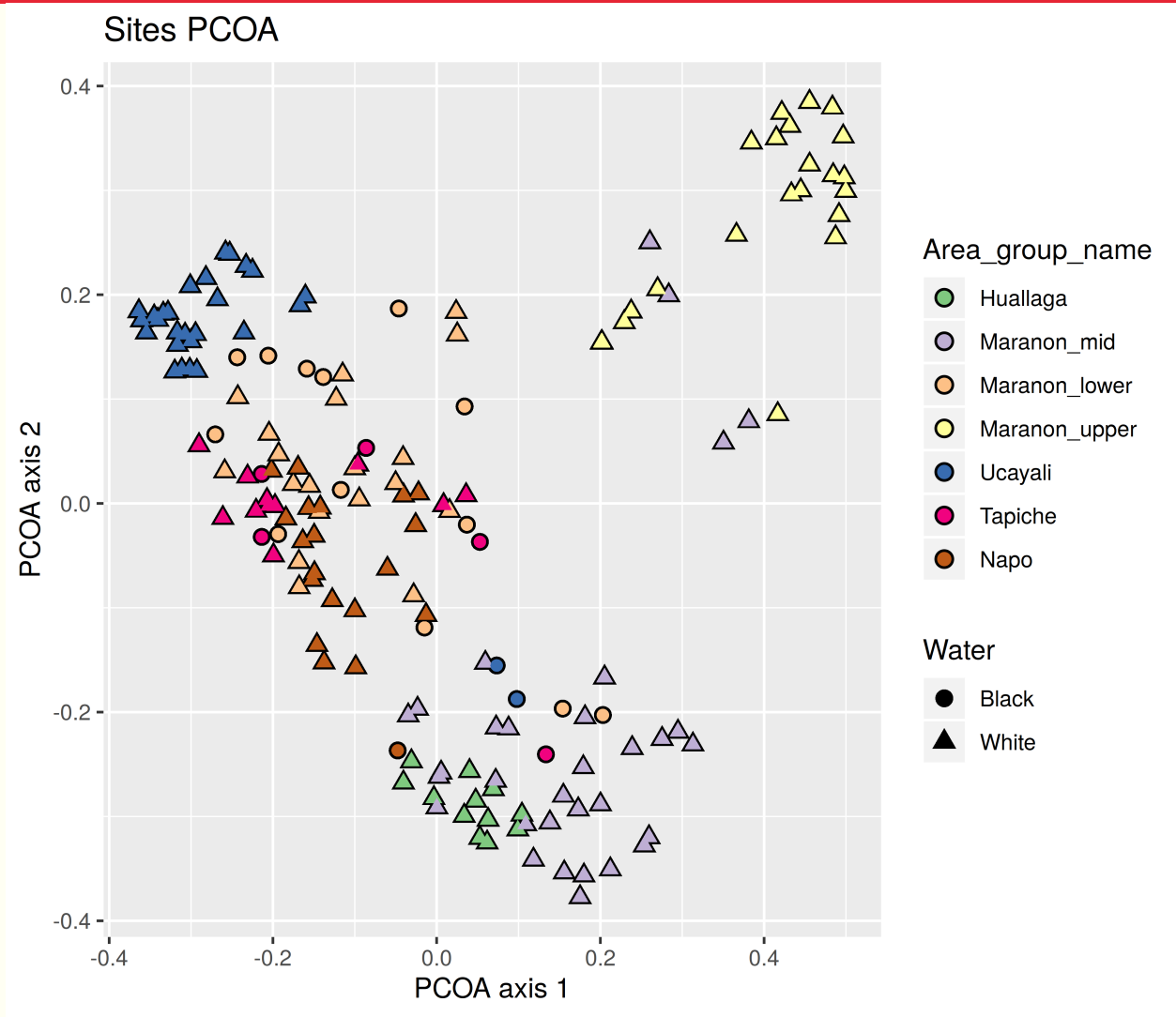
- Maximum Dissimilarity is a harder setting than similarity for prediction
- CSS normalisation and Log transformation produce the best feature set. Ordination methods may not be useful.
- Only a small set of taxonomic Orders are contributing to prediction
- Between Random Forest and Logistic regression there is no significant difference in predictive ability.
- Bayesian methods are promising but slow

Ordination Plots



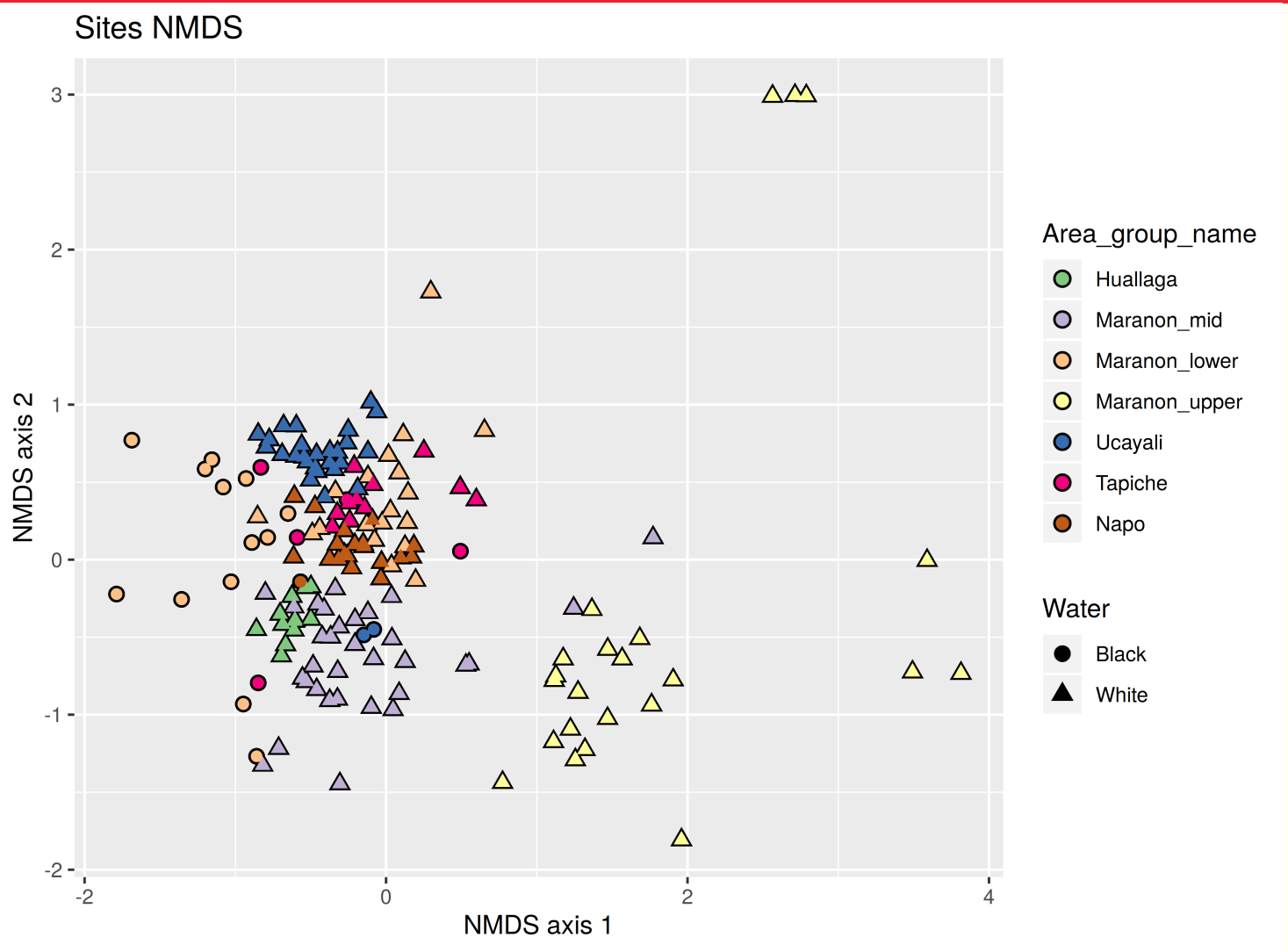
PCA

Ordination Plots



PCoA

Ordination Plots



NMDS

Benchmark

Features used	Confusion Matrix Predicted Black	Confusion Matrix Predicted White	Accuracy
Maximum Similarity			
All Labels	2.64 18.36	18.36 124.64	77.6%
Min Labels	2.75 18.25	18.25 117.75	76.8%
Maximum Dissimilarity			
All Labels	1.75 18.96	19.25 124.04	76.7%
Min Labels	1.83 18.83	19.17 117.17	75.8%

Maximally Similar LR

Features used	Confusion Matrix Predicted Black	Confusion Matrix Predicted White	Accuracy
OTU	19 1	2 142	98.17%
OTU LOW	19 0	2 143	98.78%
OTU CSS	18 0	3 143	98.17%
OTU Min CSS	18 0	3 136	98.09%
OTU CSS LOG	19 1	2 142	98.17%
PCoA Bray-Curtis	16 3	5 140	95.12%
PCoA Bray-Curtis CSS	16 1	5 142	96.34%

Maximally Similar RF

Features used	Confusion Matrix Predicted Black	Confusion Matrix Predicted White	Accuracy
OTU	17 3	4 140	95.73%
OTU LOW	18 3	3 140	96.34%
OTU CSS	18 3	3 140	96.34%
OTU Min CSS	18 3	3 133	96.18%
OTU CSS LOG	19 3	2 140	96.95%
PCoA Bray-Curtis	6 4	15 139	88.41%
PCoA Bray-Curtis CSS	4 0	17 143	89.63%

Maximally Dissimilar LR

Features used	Confusion Matrix Predicted Black	Confusion Matrix Predicted White	Accuracy
OTU	5 18	16 125	79.27%
OTU LOW	5 17	16 126	79.88%
OTU CSS	7 14	14 129	82.93%
OTU Min CSS	3 14	18 122	79.61%
OTU CSS LOG	5 4	16 139	87.80%
PCoA Bray-Curtis	7 8	14 135	86.59%
PCoA Bray-Curtis CSS	9 8	12 135	87.80%

Maximally Dissimilar RF

Features used	Confusion Matrix Predicted Black	Confusion Matrix Predicted White	Accuracy
OTU	4 10	17 133	83.50%
OTU LOW	2 10	19 133	82.32%
OTU CSS	7 2	14 141	90.24%
OTU Min CSS	7 1	14 135	90.45%
OTU CSS LOG	7 2	14 141	90.24%
PCoA Bray-Curtis	0 1	21 142	86.59%
PCoA Bray-Curtis CSS	0 5	21 138	84.15%