High Dimensional Data Analysis Final Project

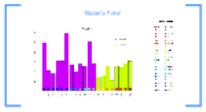
THE END

THANK YOU FOR LISTENING

MicroMass DataSet

Mass Spectrometry

- 1300 spectral variables
- 571 observations of bacterial panels
- 20 species of bacteria among 9 genera



Data Preprocessing

Missing value checking

Zero value checking

Standardized

Data Visualization

Minkowski distance heatmap

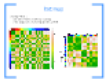
Correlation-based similarity heatmap

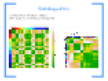
Dimension Reduction

ISOmap for 3-dimensional space

Shrinkage PCA for 3-dimensional space

MDS for 3-dimensional space



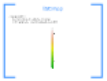


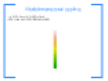
Performance of higher dimension selected

Shrinkage PCA for 30-dimension

ISOmap for 30-dimension

MDS for 30-dimension





Discussing of Optimization

The performance about different numbers of dimension selected for models

Classification

MicroMass DataSet

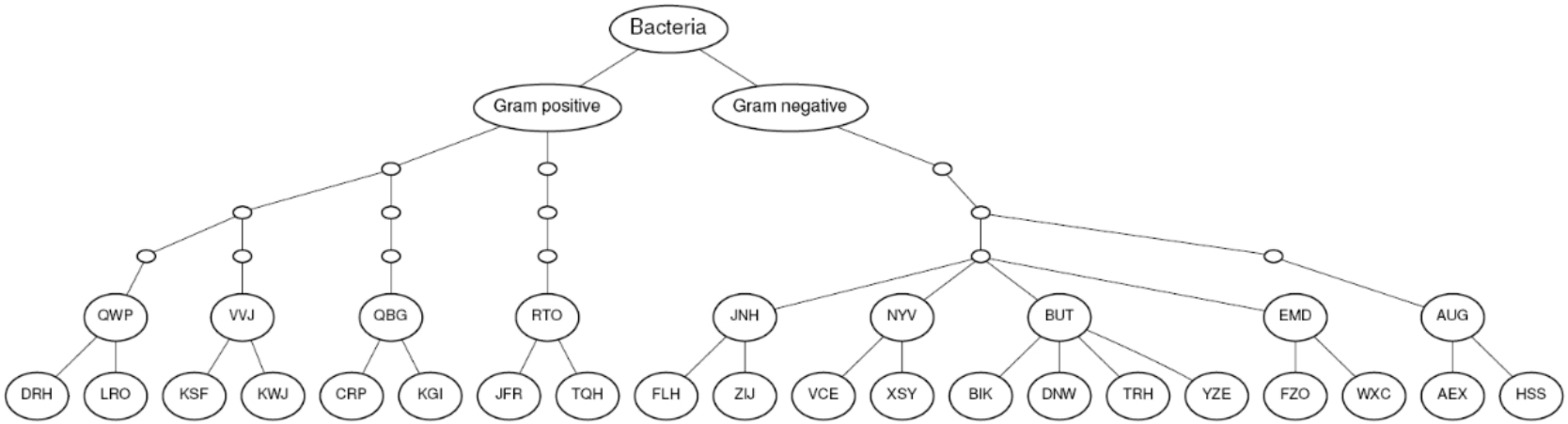
Mass Spectrometry

1300 spectral variables

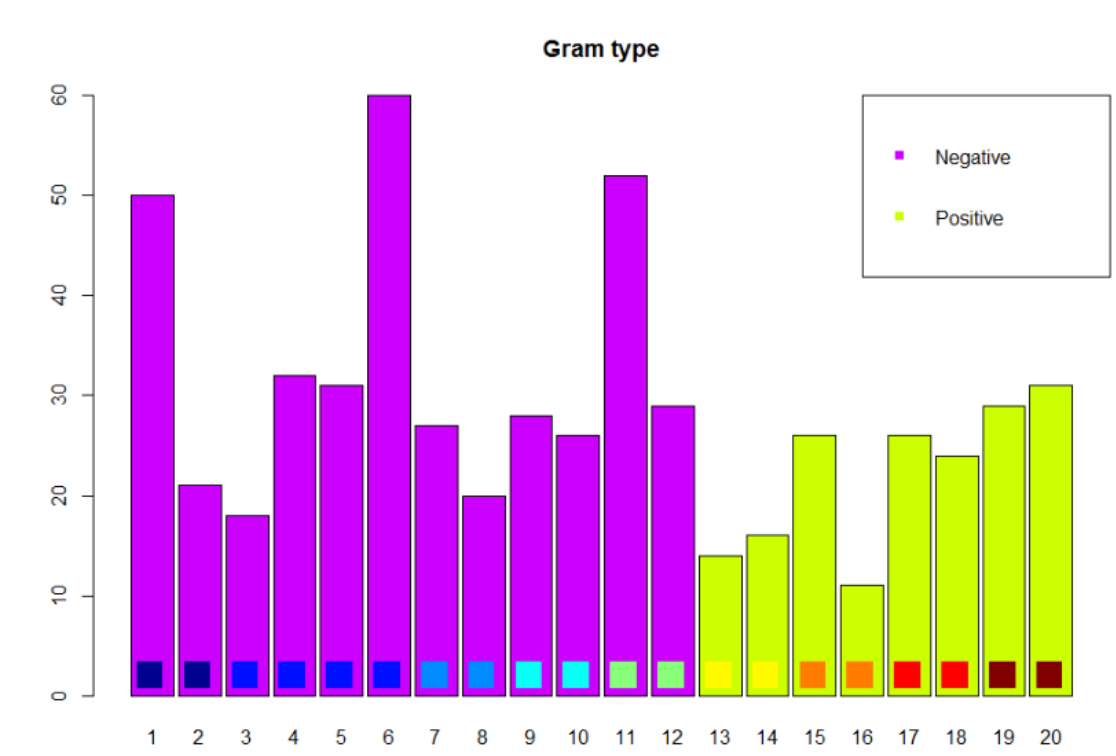
571 observations of bacterial panels

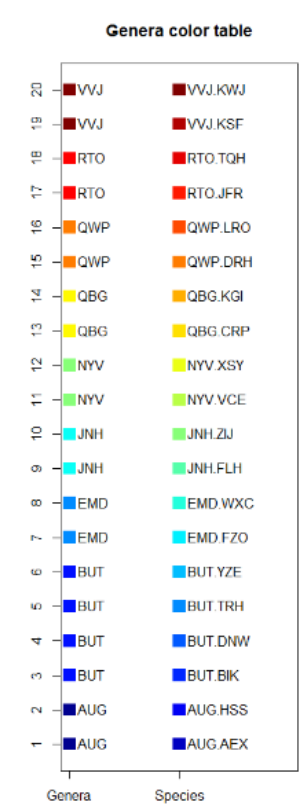
20 species of bacteria among **9** genera

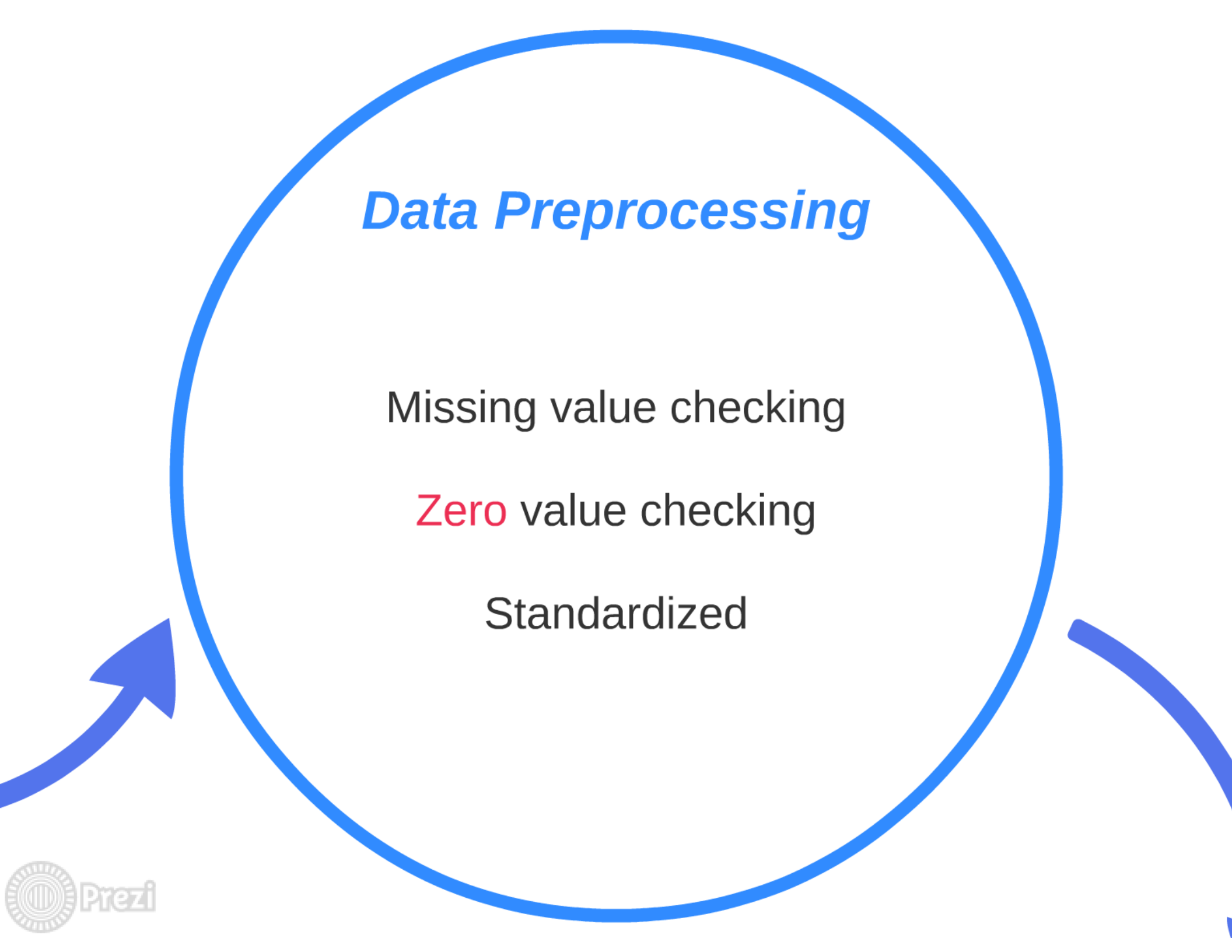
Bacteria Panel



Bacteria Panel






Data Preprocessing
Missing value checking
Zero value checking
Standardized

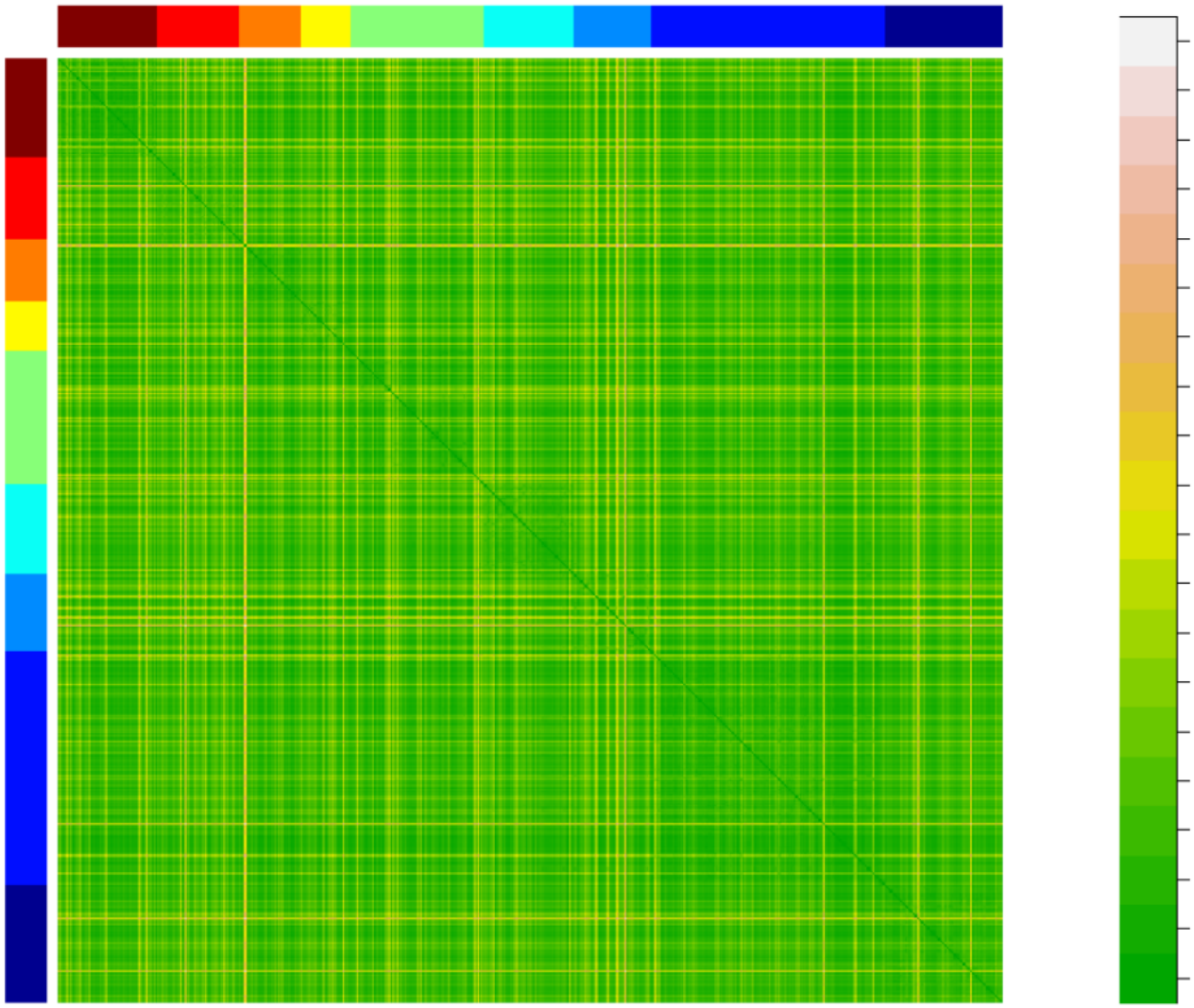
Data Visualization

Minkowski distance heatmap

Correlation-bsed similarity heatmap

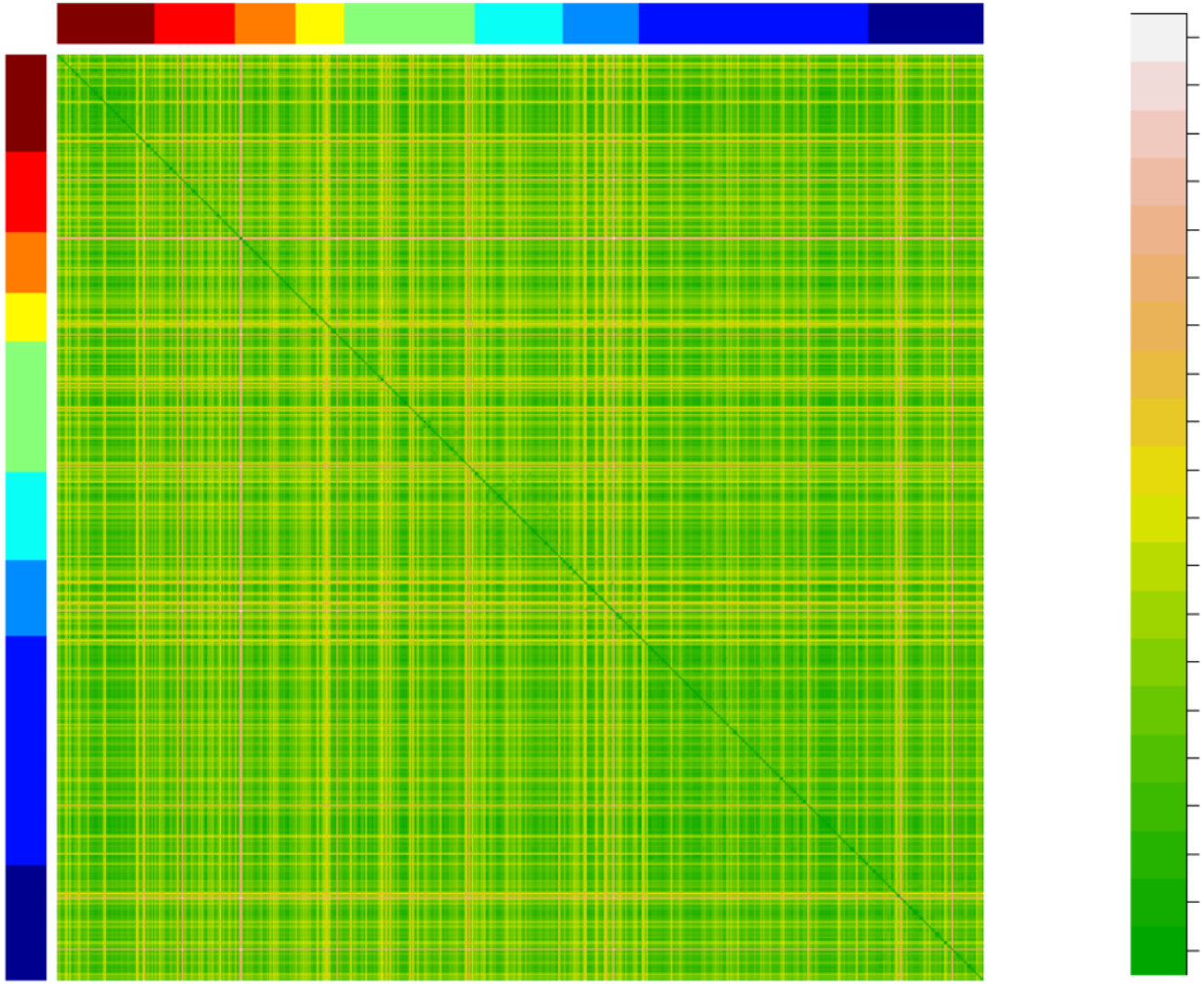
Minkowski, L=1
order by genera

Manhattan distance heatmap



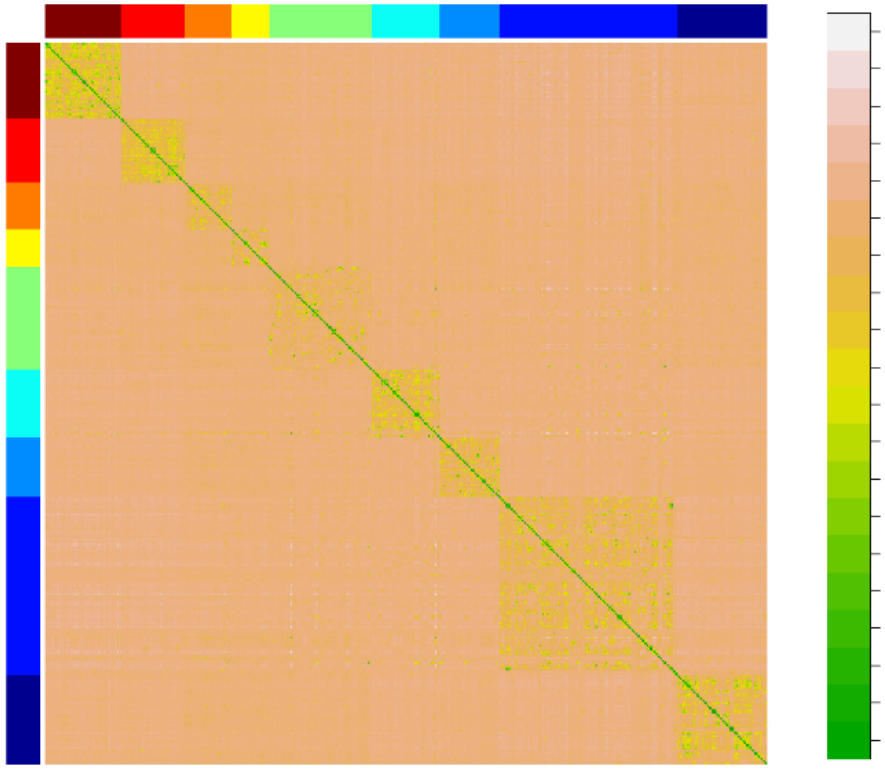
Minkowski, L=2
order by genera

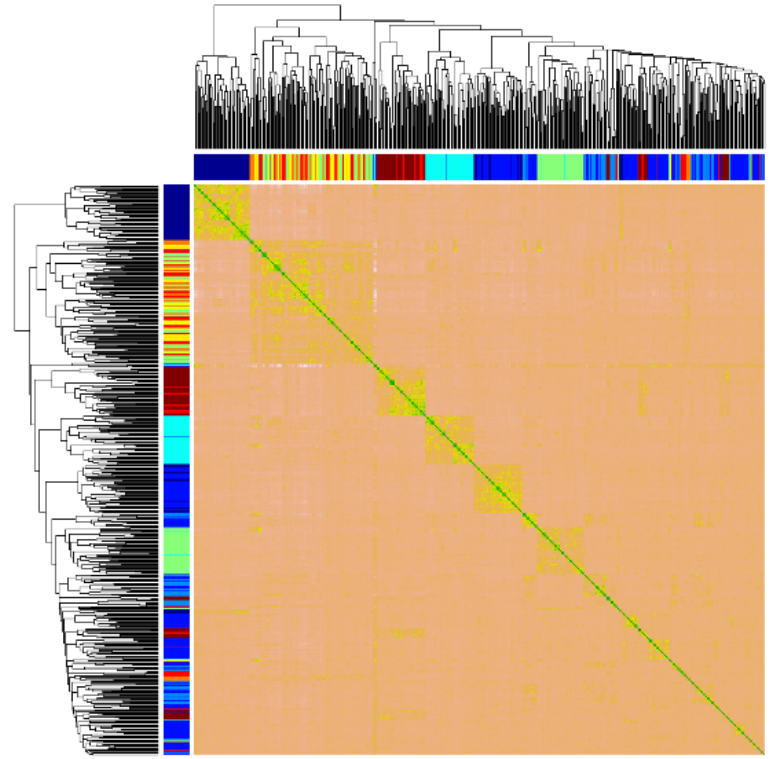
Euclidean distance heatmap



Correlation-based similarity heatmap

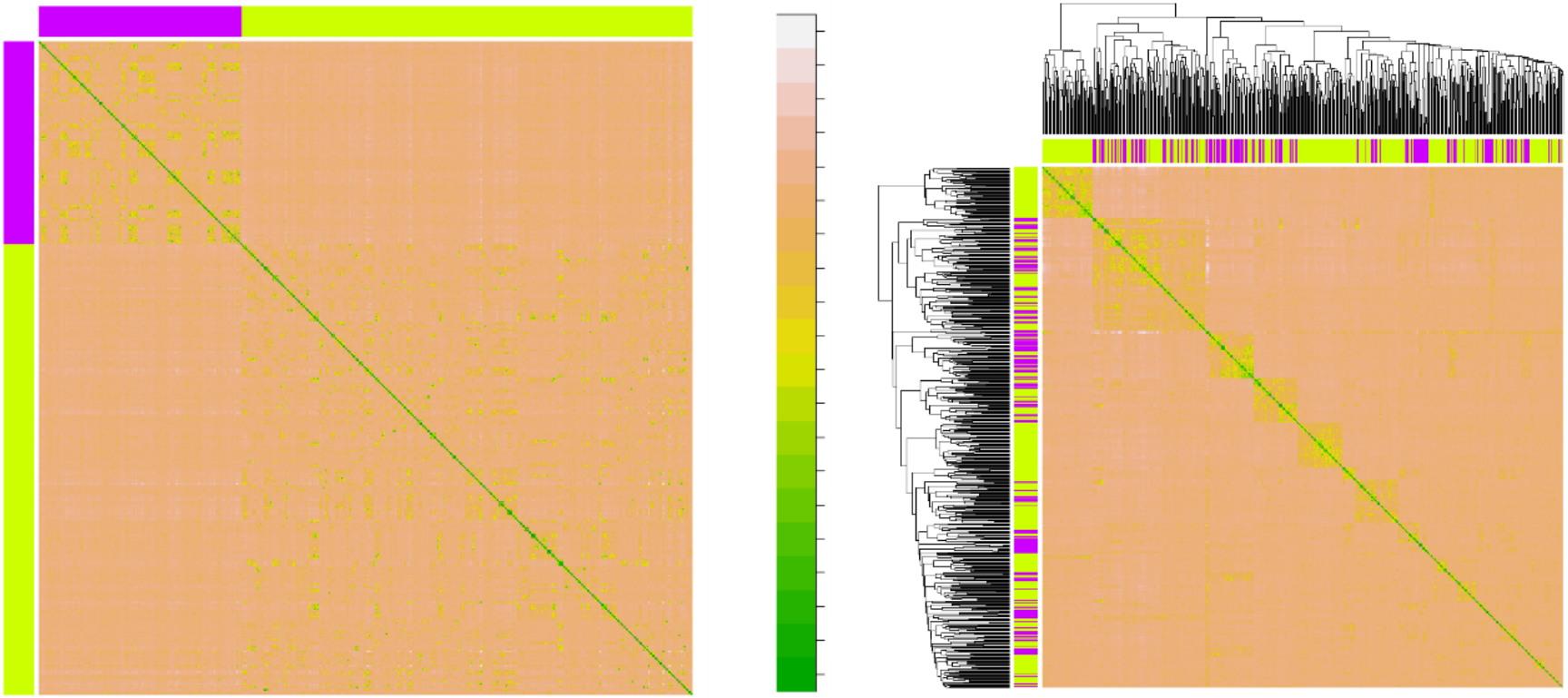
order by genera, clustering by average-link





Correlation-based similarity heatmap

order by gram type, clustering by average-link

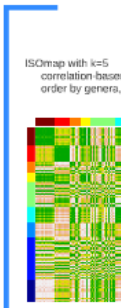


Dimension Reduction

ISOmap for 3-dimensional space

Shrinkage PCA for 3-dimensional space

MDS for 3-dimensional space

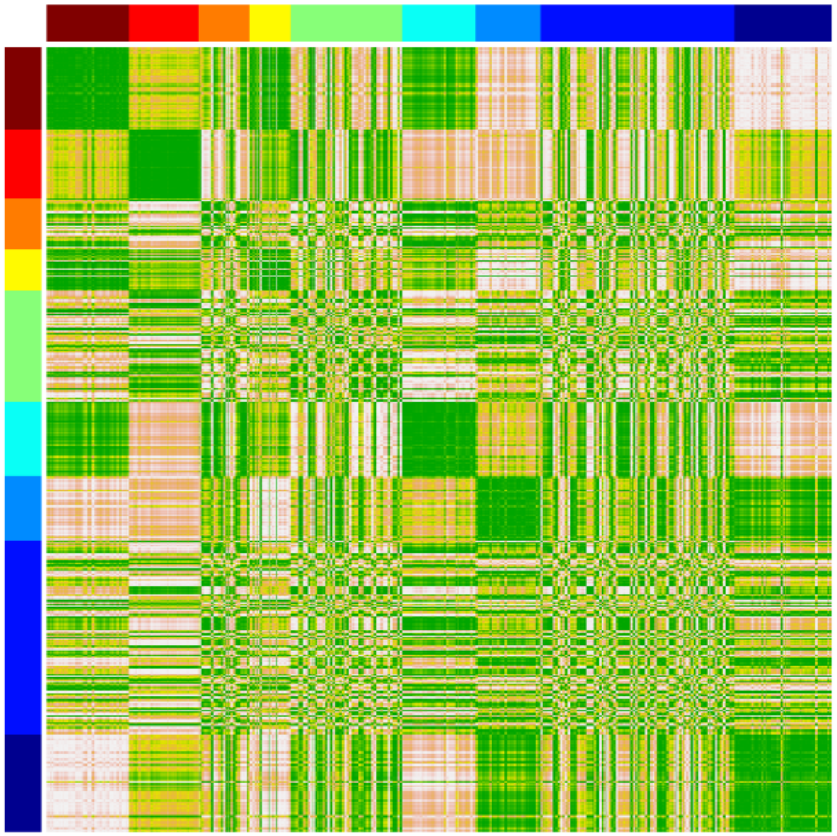


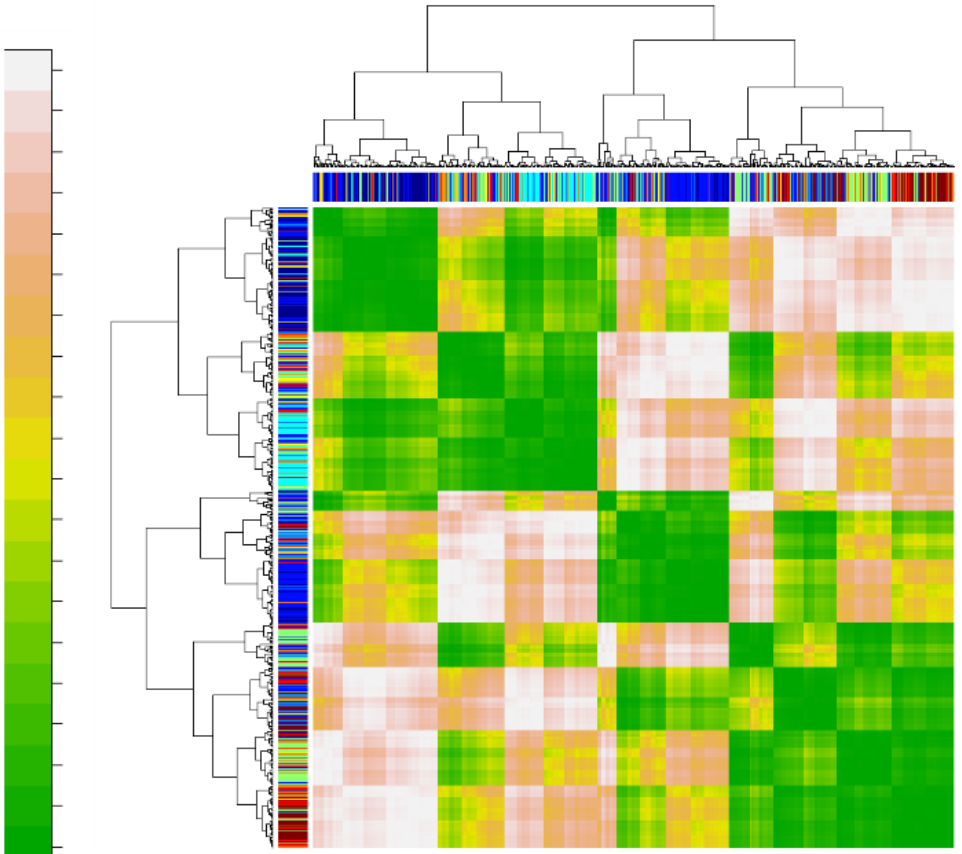
ISOmap

ISOmap with k=5

correlation-based similarity heatmap

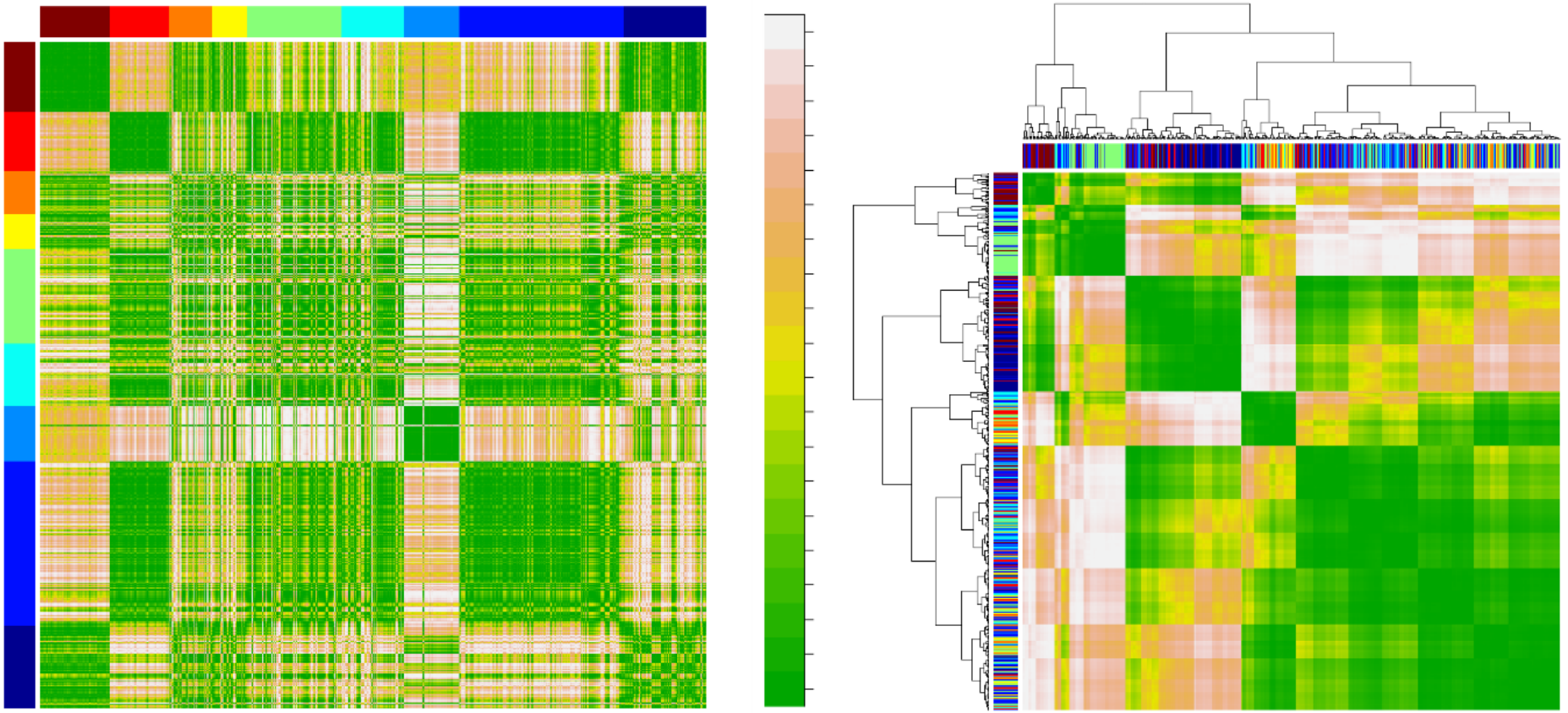
order by genera, clustering by average-link





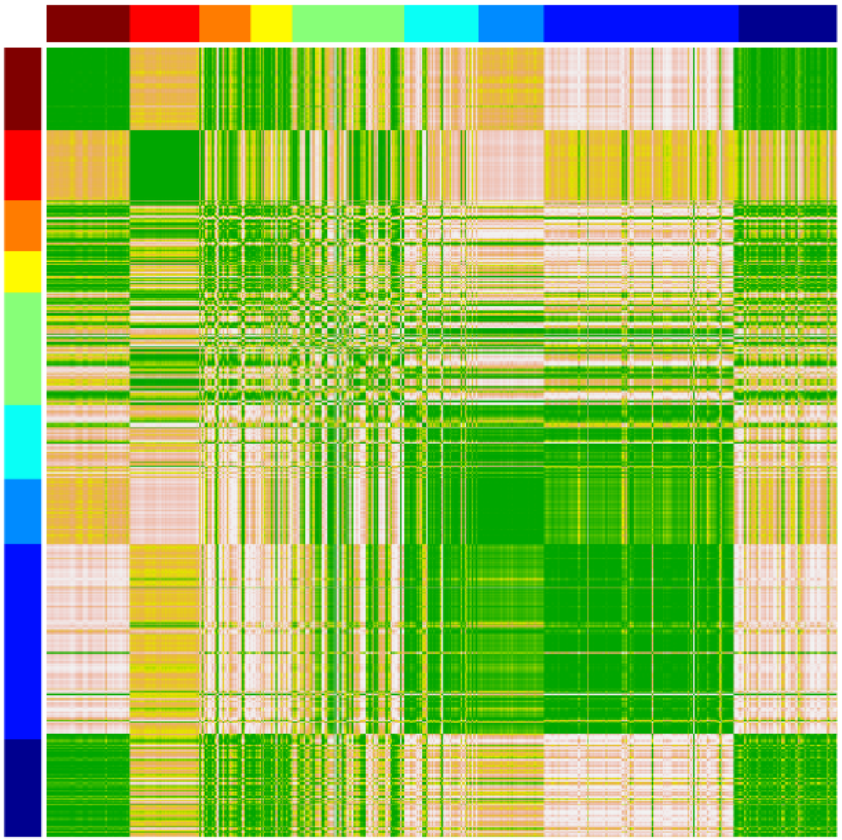
Shrinkage PCA

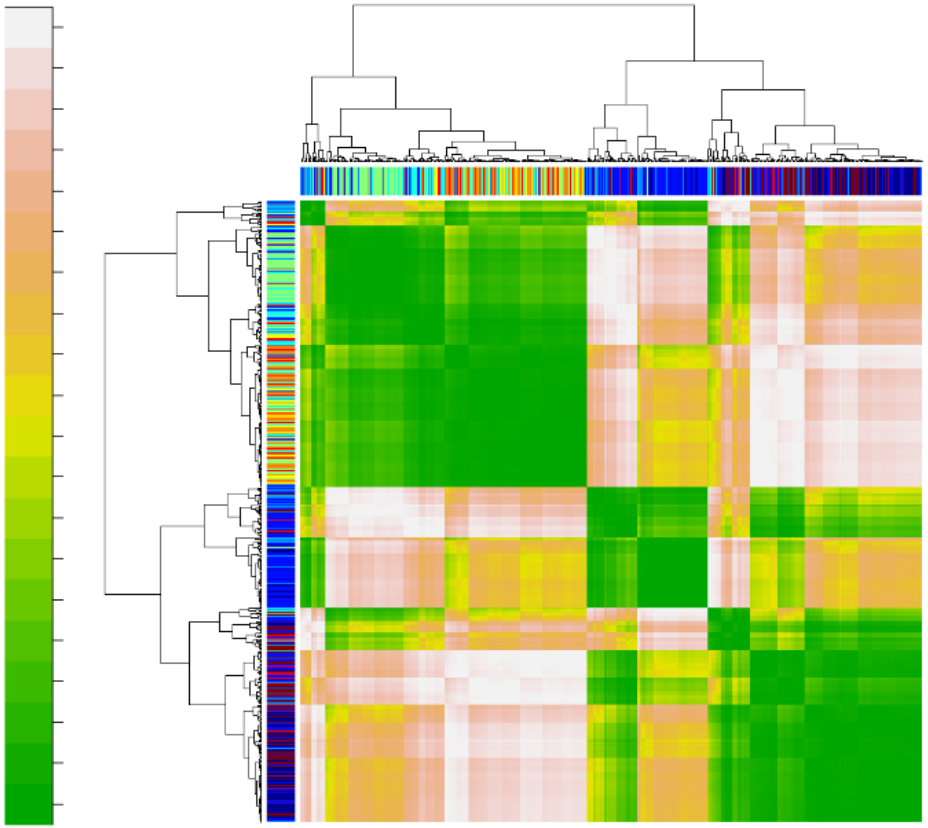
correlation-based similarity heatmap
order by genera, clustering by average-link



Multidimensional scaling

correlation-based similarity heatmap
order by genera, clustering by average-link



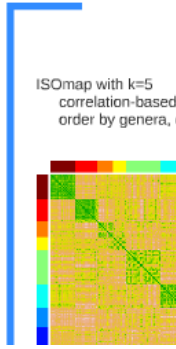


Performance of higher dimension selected

Shrinkage PCA for 30-dimension

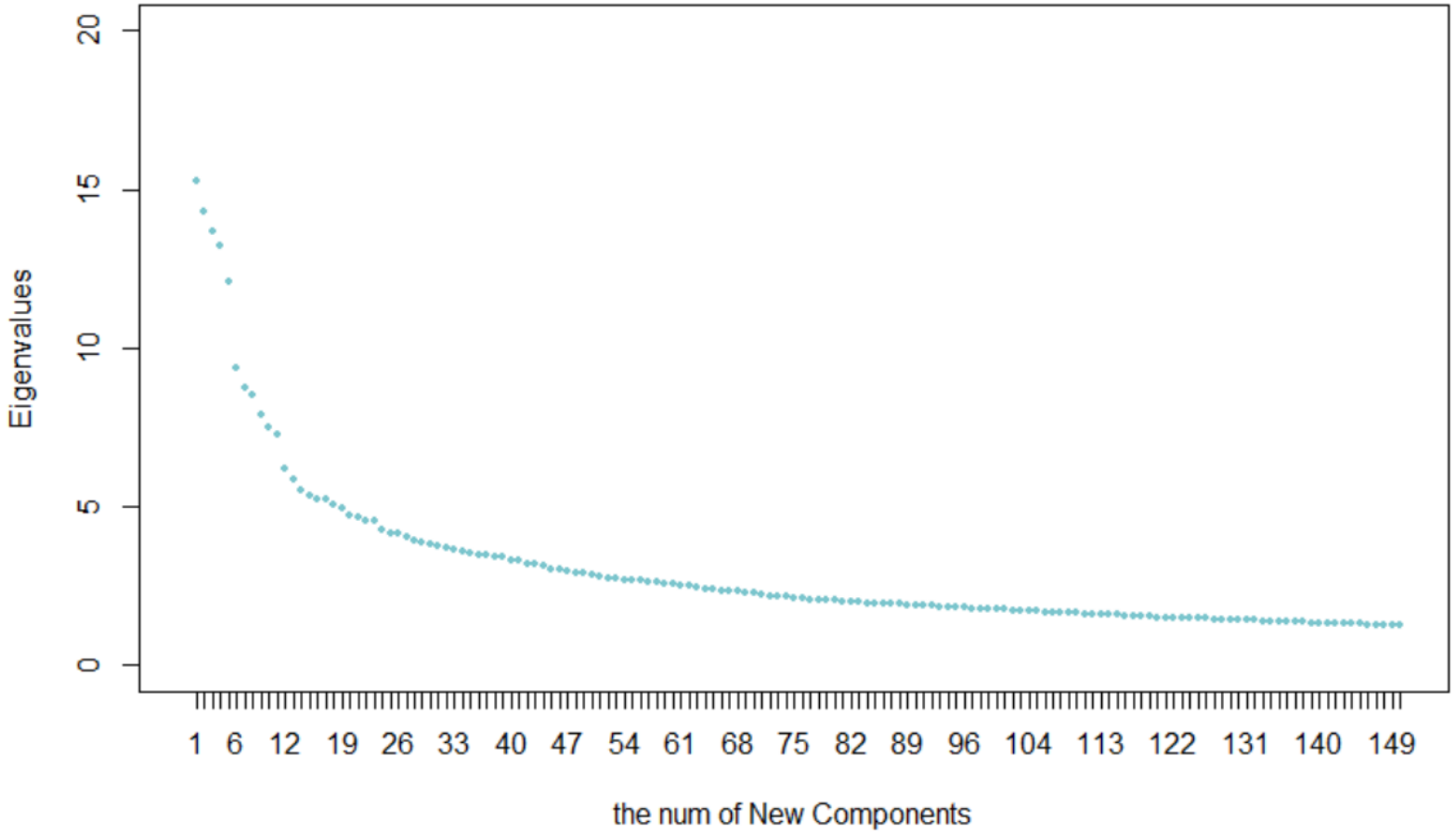
ISOmap for 30-dimension

MDS for 30-dimension



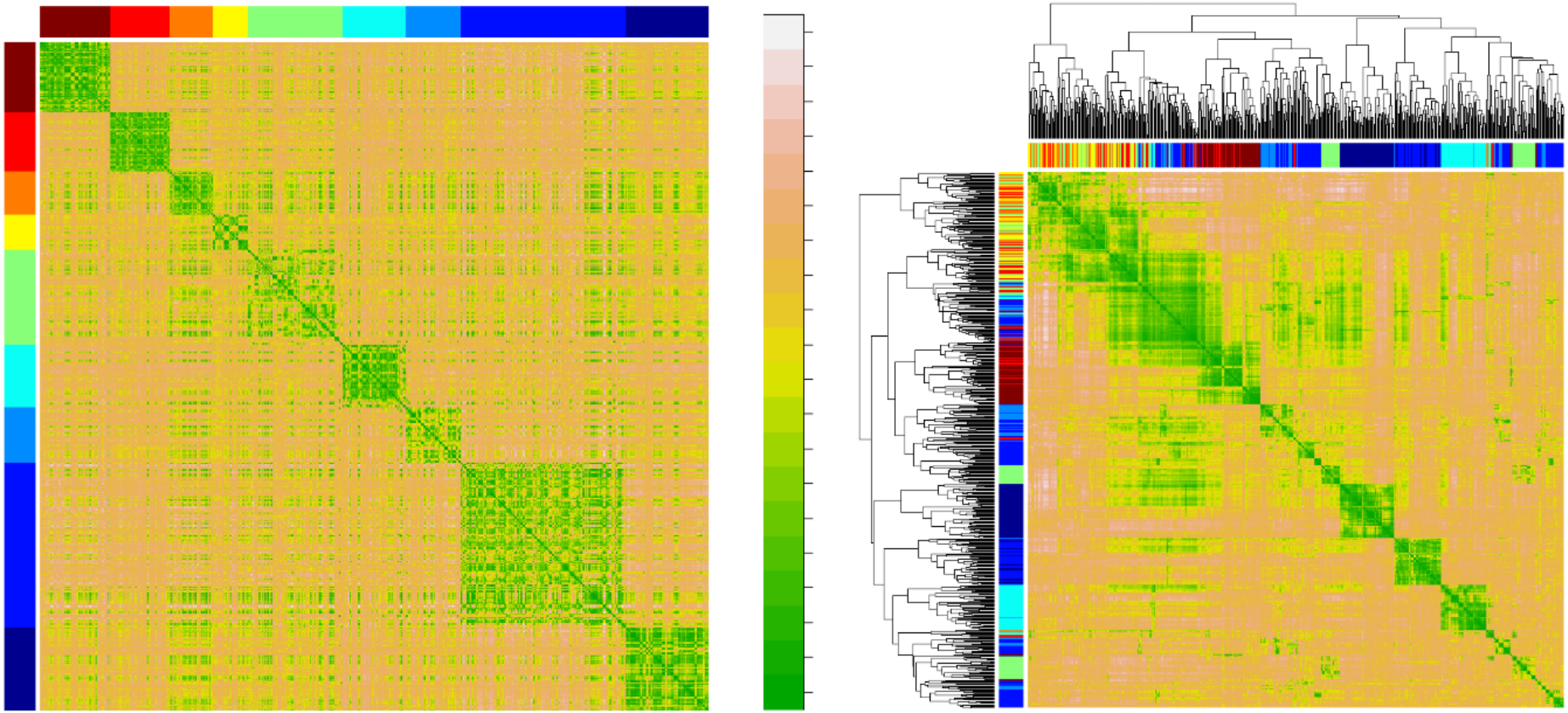
Shrinkage PCA

eigenvalues plot



Shrinkage PCA

correlation-based similarity heatmap
order by genera, clustering by average-link

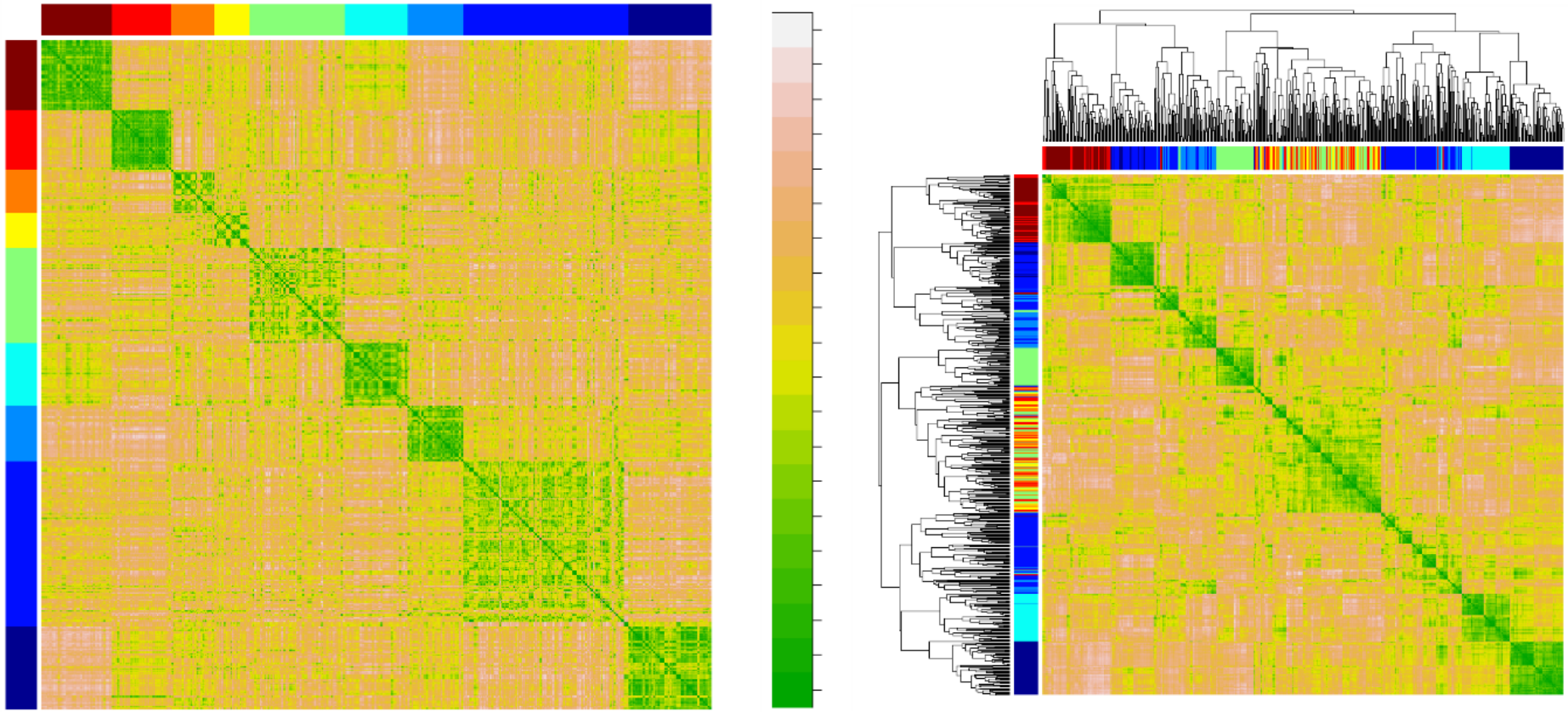


ISOmap

ISOmap with k=5

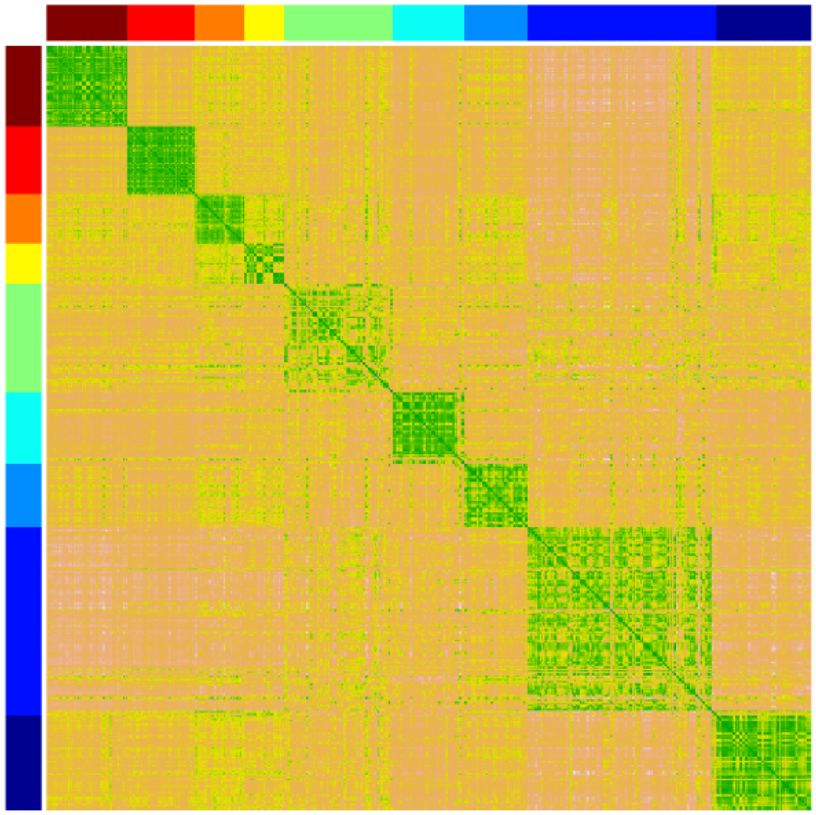
correlation-based similarity heatmap

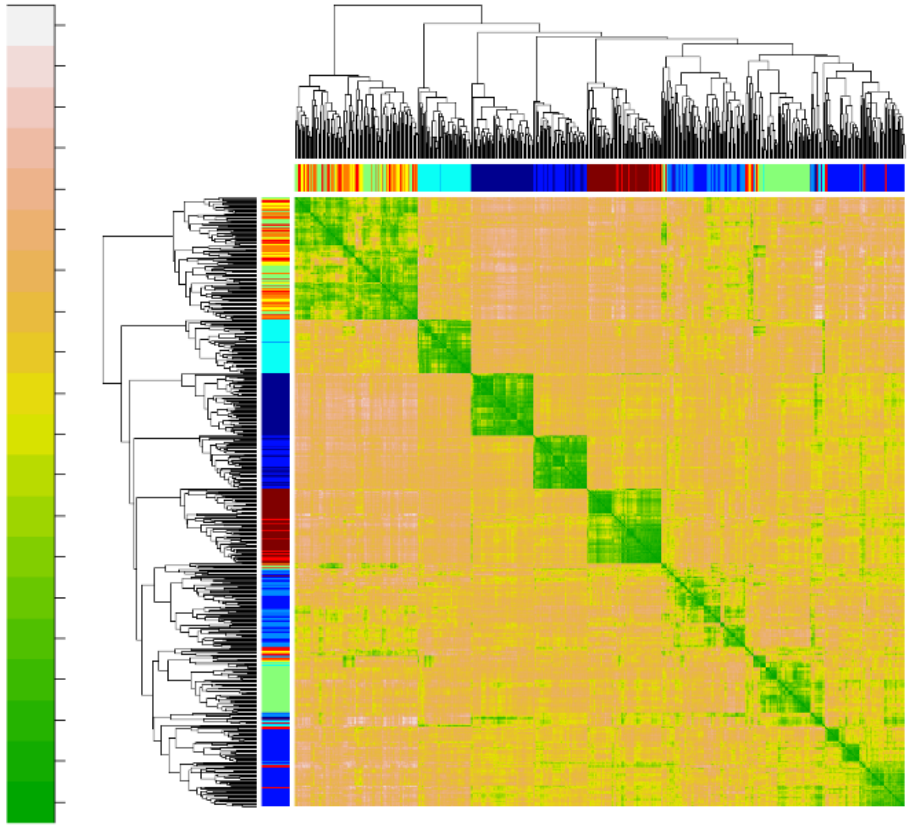
order by genera, clustering by average-link



Multidimensional scaling

correlation-based similarity heatmap
order by genera, clustering by average-link





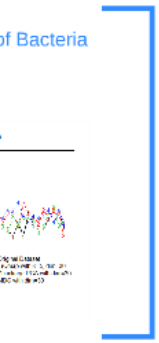
Classification

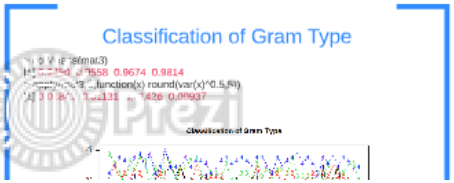
Support vector machine with kernel = linear

Subsampling the training dataset and the testing dataset with the ratio 2 : 1

The accuracy of prediction is obtained from processing SVM for 50 times

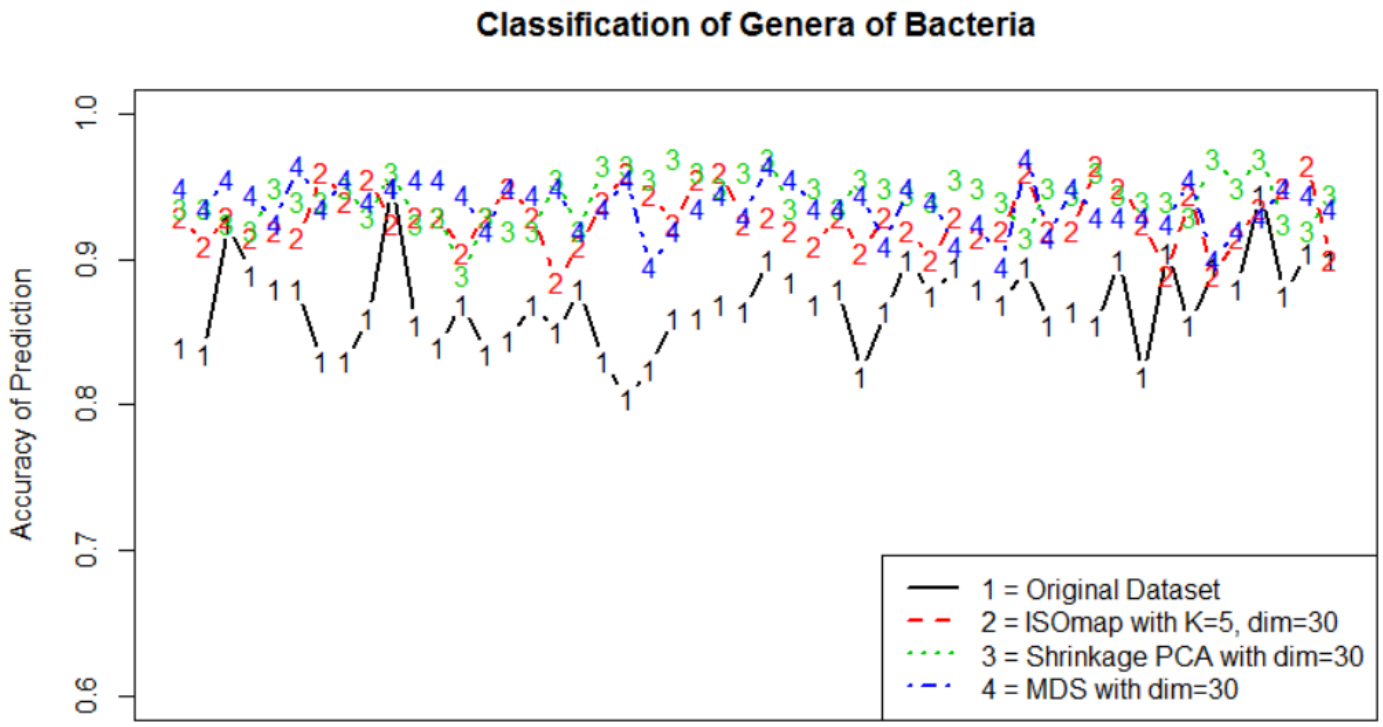
Stability of the classifying performance of SVM





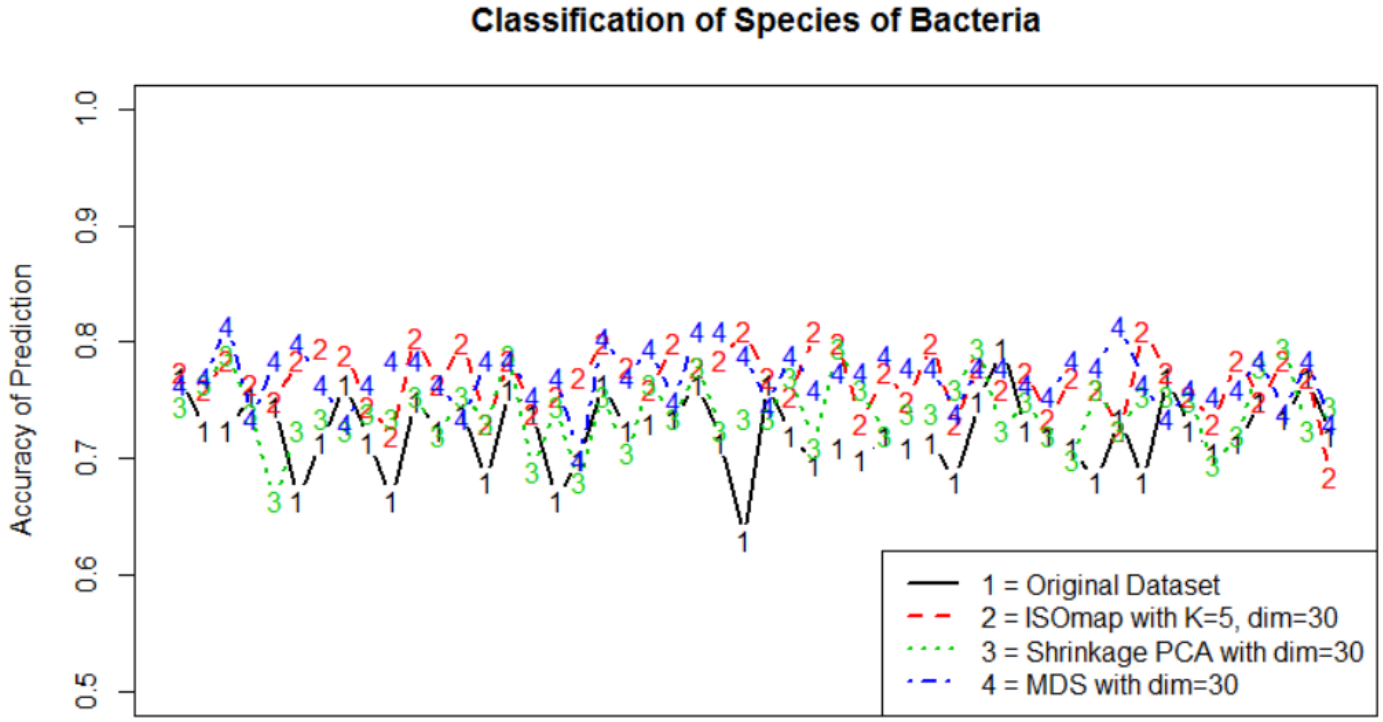
Classification of Genera of Bacteria

```
> colMeans(mat1)
[1] 0.8694  0.9284  0.9426  0.9372
> apply(mat1,2,function(x) round(var(x)^0.5,5))
[1] 0.03092  0.02026  0.01703  0.01762
```



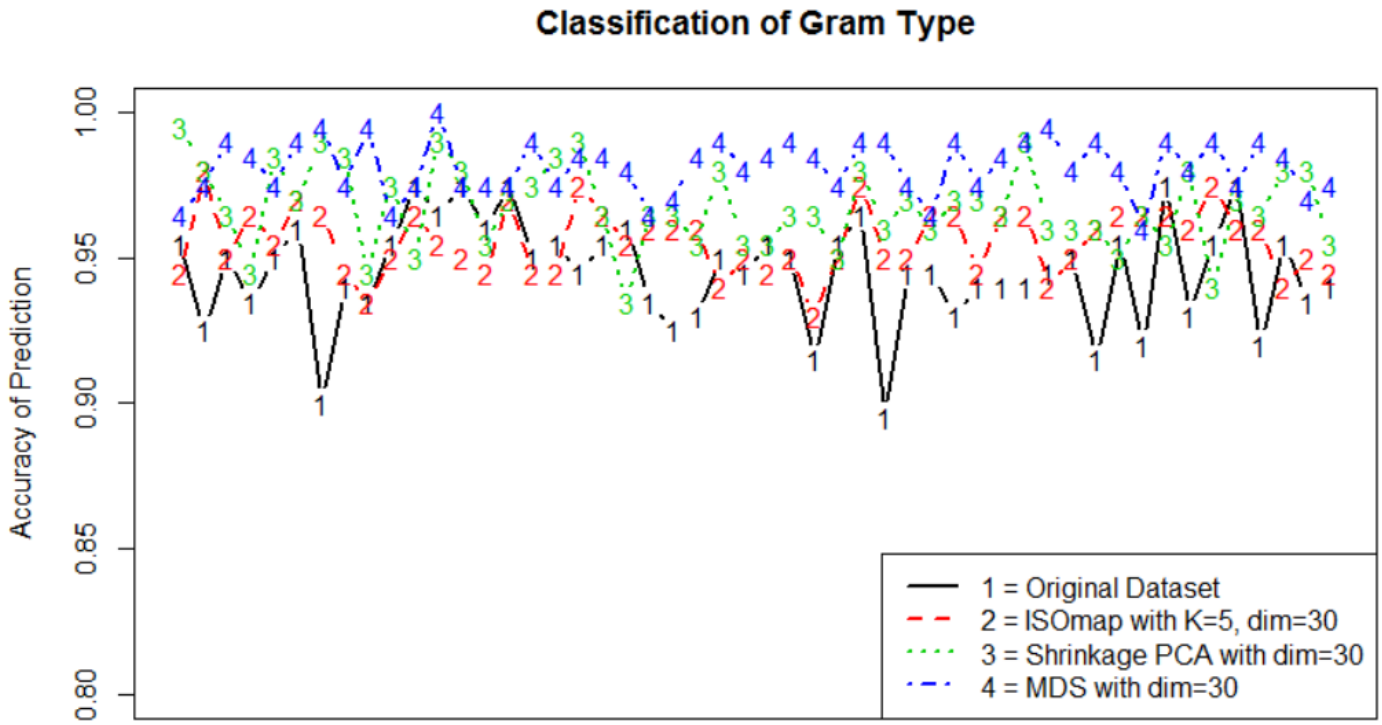
Classification of Species of Bacteria

```
> colMeans(mat2)
[1] 0.7234  0.7681  0.7411  0.7707
> apply(mat2,2,function(x) round(var(x)^0.5,5))
[1] 0.03325  0.02723  0.02949  0.02443
```



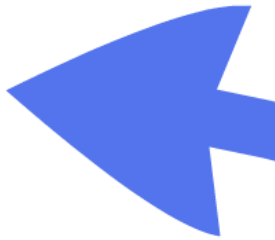
Classification of Gram Type

```
> colMeans(mat3)
[1] 0.9450  0.9558  0.9674  0.9814
> apply(mat3,2,function(x) round(var(x)^0.5,5))
[1] 0.01841  0.01131  0.01426  0.00937
```

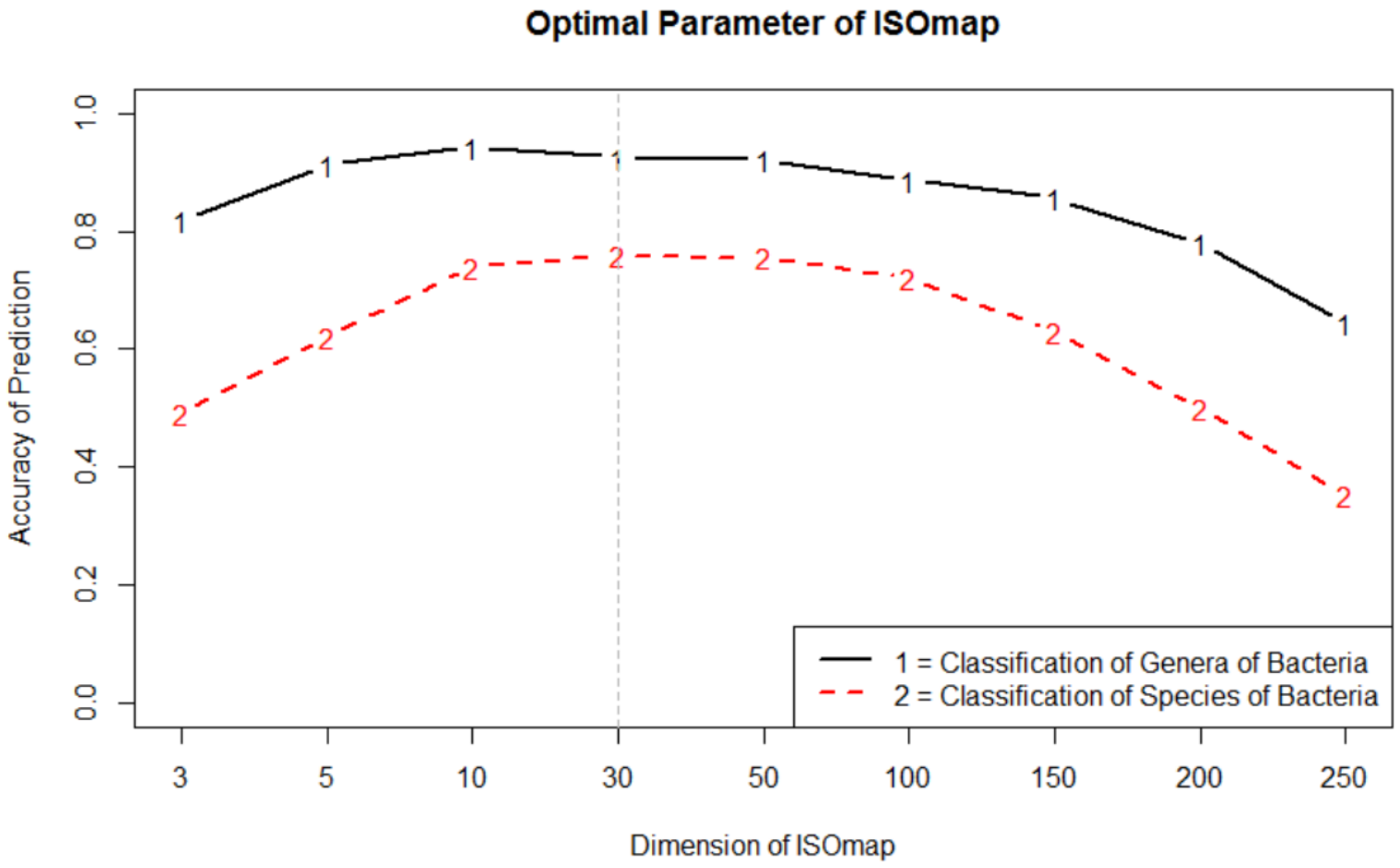


Discussing of Optimization

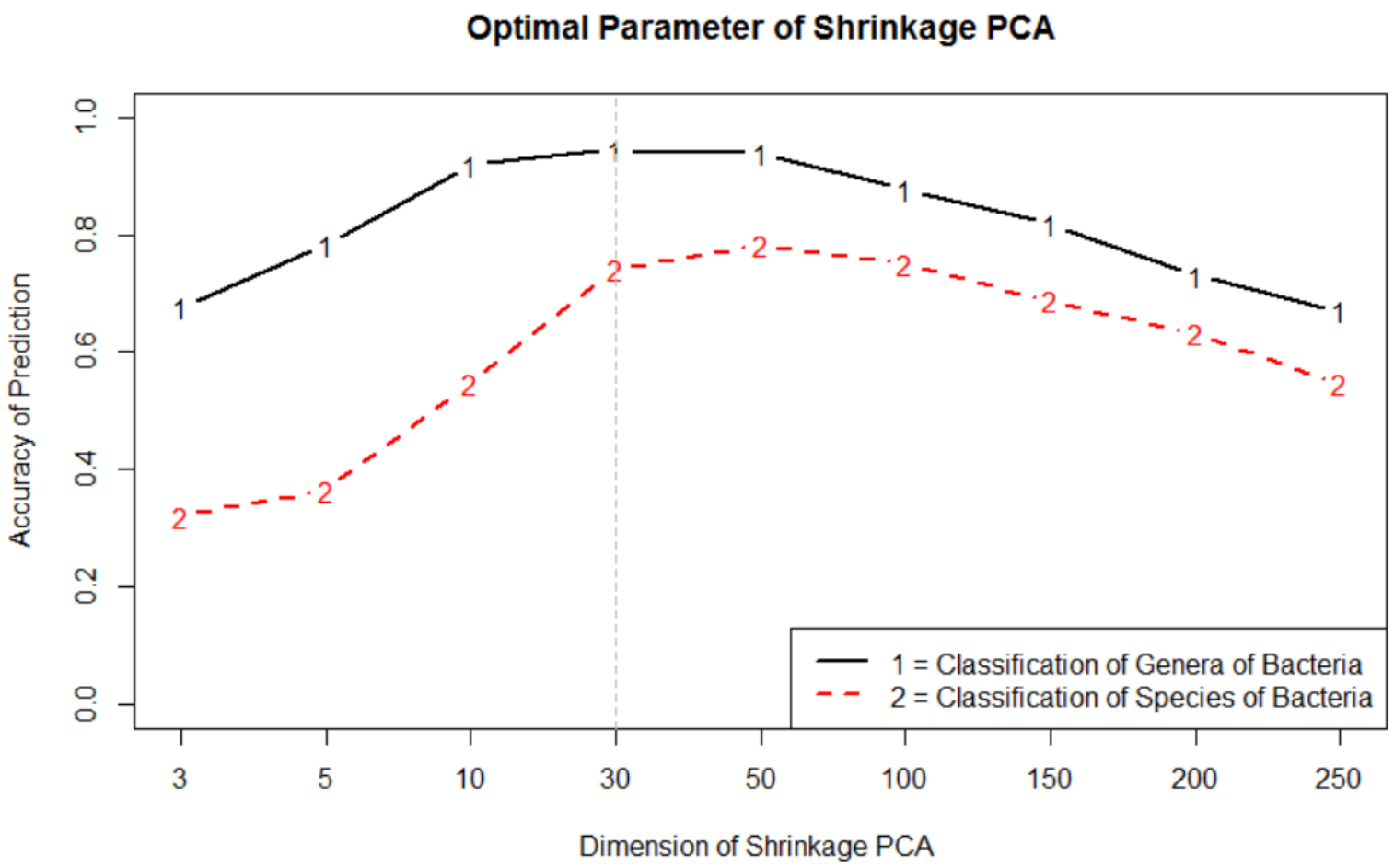
The performance about different numbers of dimension selected for models



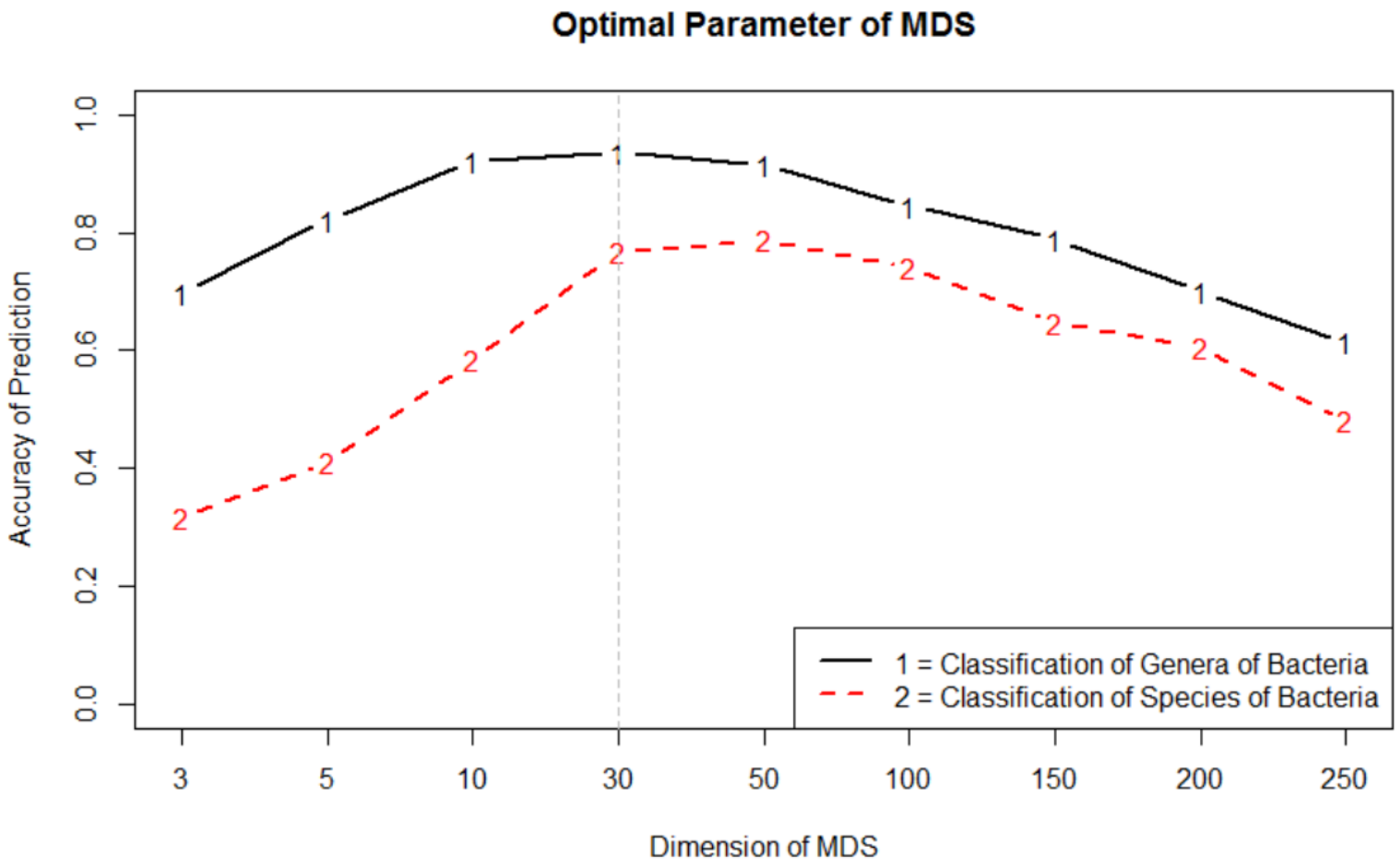
ISOmap


Optimal Parameter of ISOmap
Accuracy of Prediction
0.0
0.2
0.4
0.6
0.8
1.0
3
5
10
30
50
100
150
200
250
Dimension of ISOmap
1 = Classification of Genera of Bacteria
2 = Classification of Species of Bacteria

Shrinkage PCA



Multidimensional scaling



THE END

THANK YOU FOR LISTENING