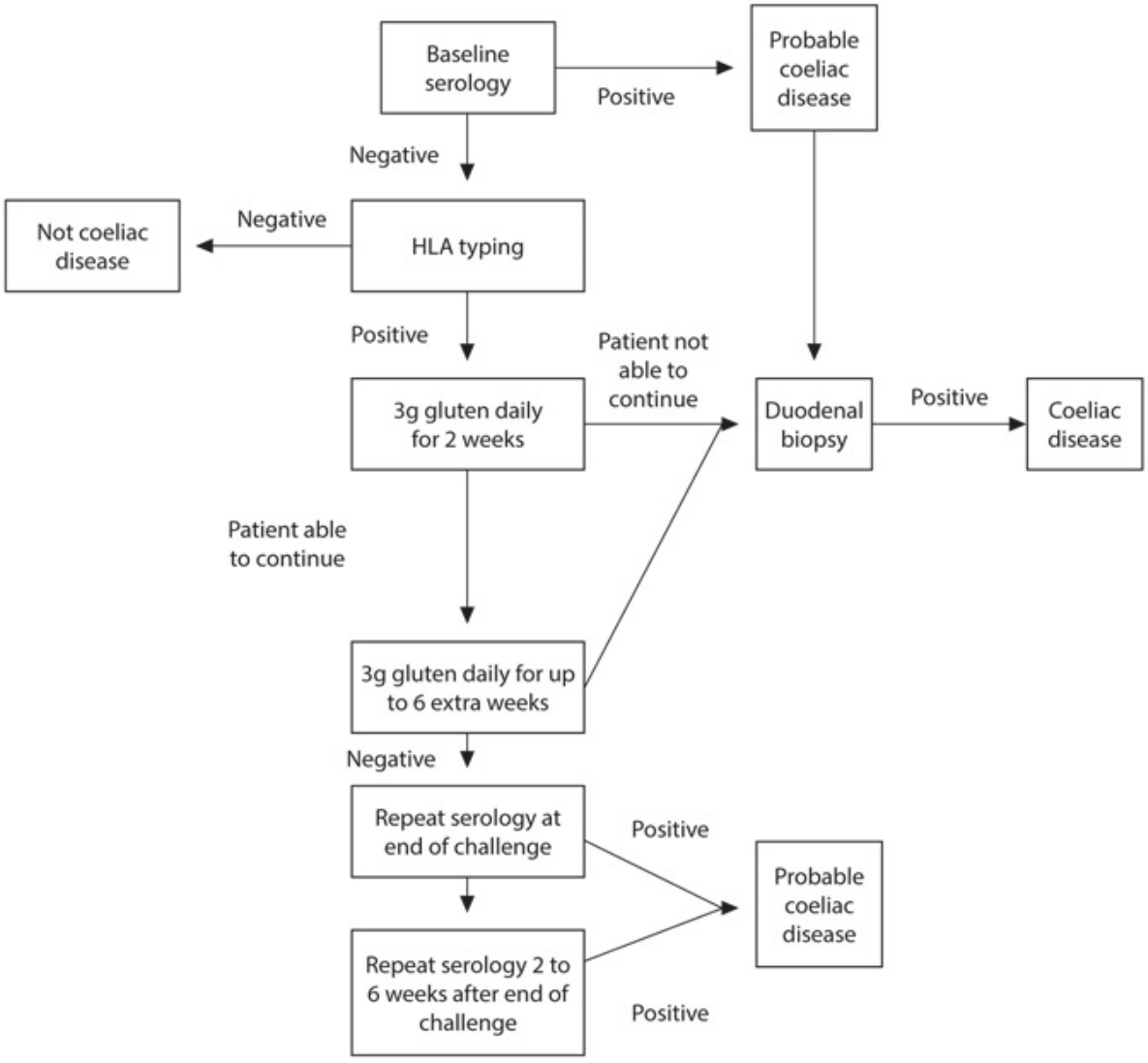
Baseline serology
Positive
Probable coeliac disease
Negative
Not coeliac disease
Negative
HLA typing
Positive
Patient not able to continue
3g gluten daily for 2 weeks
Duodenal biopsy
Positive
Coeliac disease
Patient able to continue
3g gluten daily for up to 6 extra weeks
Negative
Repeat serology at end of challenge
Positive
Probable coeliac disease
Repeat serology 2 to 6 weeks after end of challenge
Positive

Figure 1

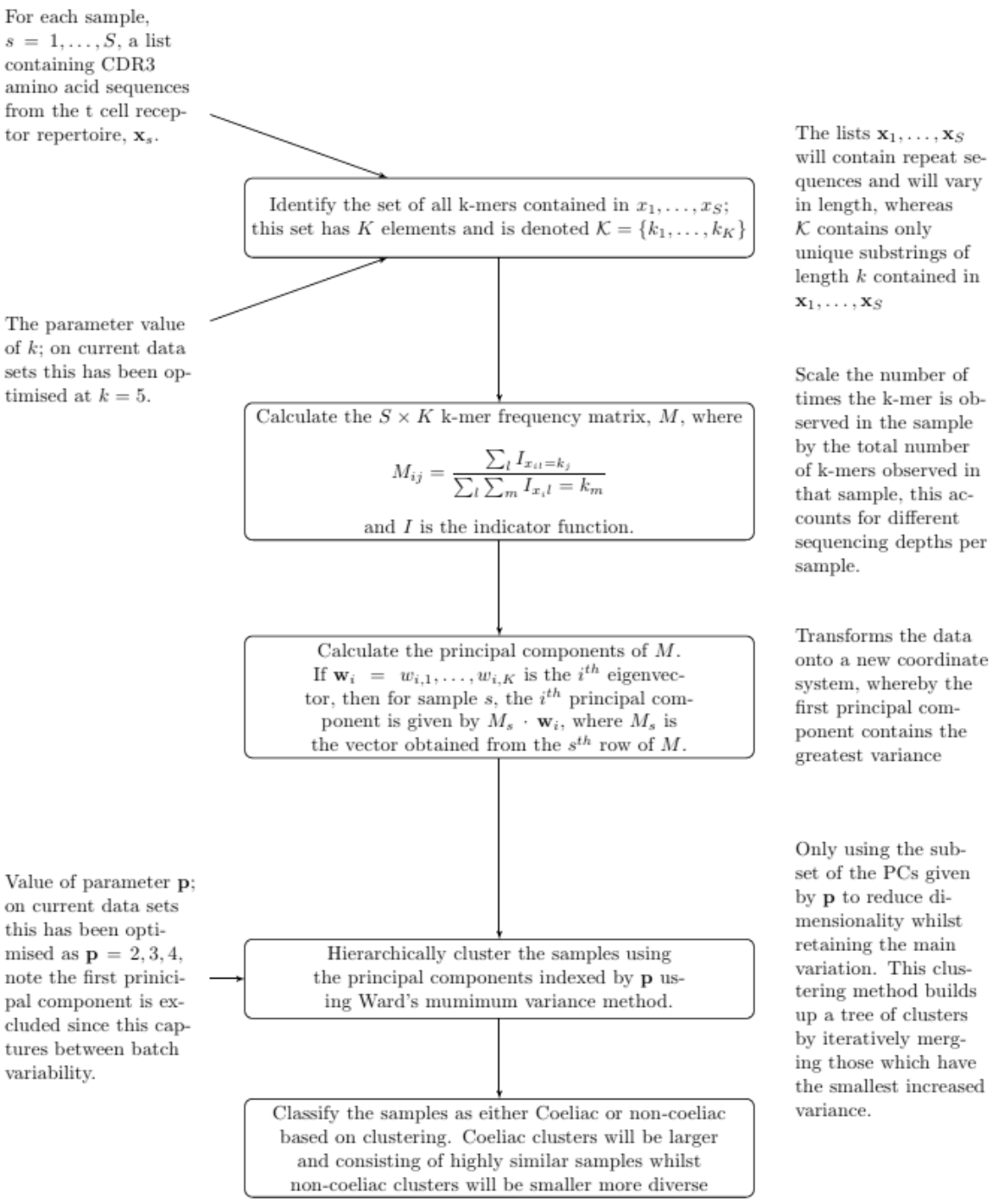
For each sample, $s = 1, \ldots, S$, a list containing CDR3 amino acid sequences from the t cell receptor repertoire, $\mathbf{x}_s$.
Identify the set of all k-mers contained in $x_1, \ldots, x_S$; this set has K elements and is denoted $\mathcal{K} = \{k_1, \ldots, k_K\}$
The lists $\mathbf{x}_1, \ldots, \mathbf{x}_S$ will contain repeat sequences and will vary in length, whereas $\mathcal{K}$ contains only unique substrings of length k contained in $\mathbf{x}_1, \ldots, \mathbf{x}_S$
The parameter value of k; on current data sets this has been optimised at $k = 5$.
Calculate the $S \times K$ k-mer frequency matrix, M, where
$M_{ij} = \frac{\sum_l I_{x_{il}=k_j}}{\sum_l \sum_m I_{x_i l}=k_m}$
and I is the indicator function.
Scale the number of times the k-mer is observed in the sample by the total number of k-mers observed in that sample, this accounts for different sequencing depths per sample.
Calculate the principal components of M. If $\mathbf{w}_i = w_{i,1}, \ldots, w_{i,K}$ is the i^{th} eigenvector, then for sample s, the i^{th} principal component is given by $M_s \cdot \mathbf{w}_i$, where M_s is the vector obtained from the s^{th} row of M.
Transforms the data onto a new coordinate system, whereby the first principal component contains the greatest variance
Value of parameter $\mathbf{p}$; on current data sets this has been optimised as $\mathbf{p} = 2, 3, 4$, note the first prinicipal component is excluded since this captures between batch variability.
Hierarchically cluster the samples using the principal components indexed by $\mathbf{p}$ using Ward's mumimum variance method.
Only using the subset of the PCs given by $\mathbf{p}$ to reduce dimensionality whilst retaining the main variation. This clustering method builds up a tree of clusters by iteratively merging those which have the smallest increased variance.
Classify the samples as either Coeliac or non-coeliac based on clustering. Coeliac clusters will be larger and consisting of highly similar samples whilst non-coeliac clusters will be smaller more diverse

Figure 2a

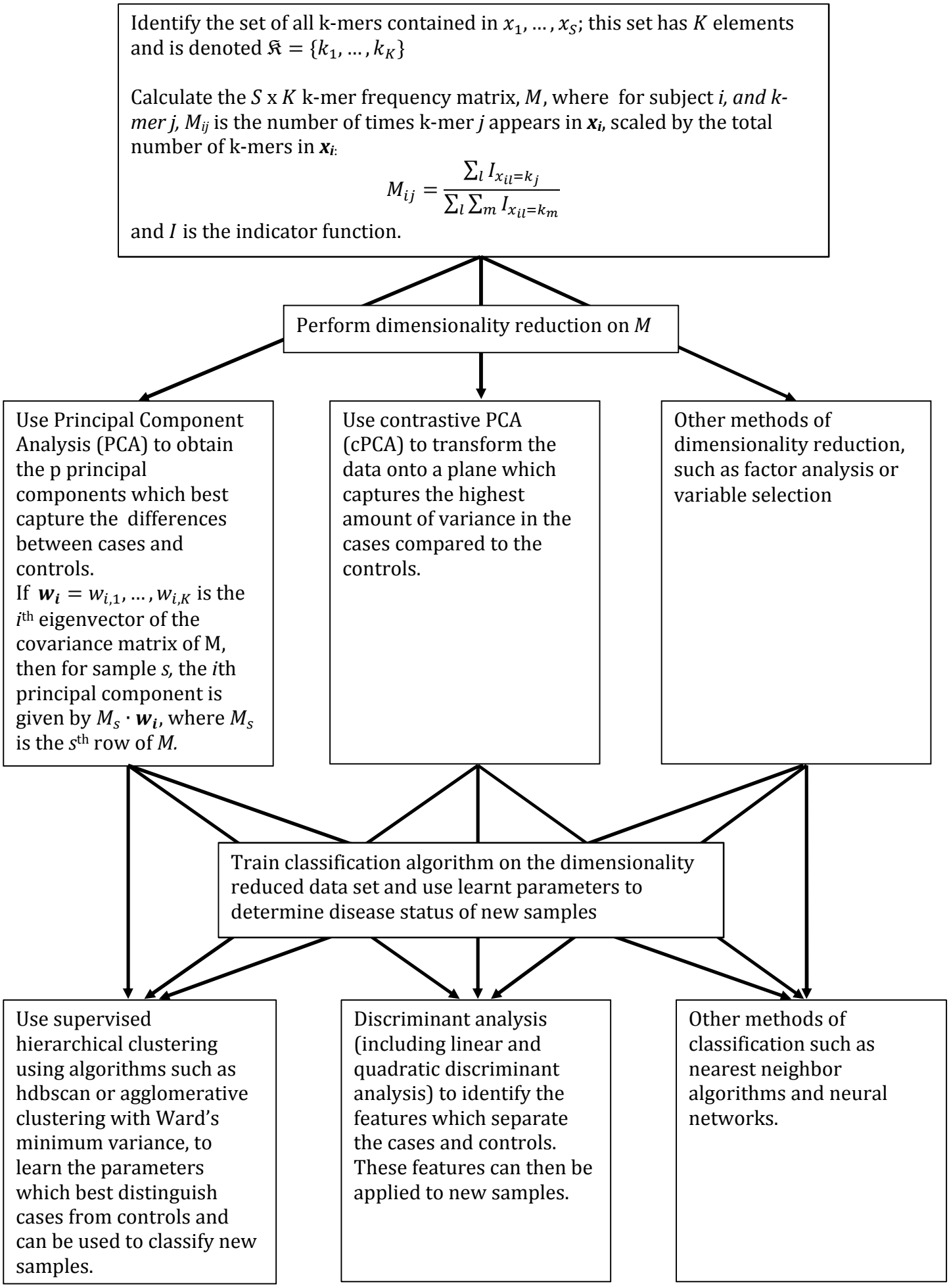
Identify the set of all k-mers contained in $x_1, \dots, x_S$; this set has K elements and is denoted $\mathfrak{K} = \{k_1, \dots, k_K\}$
Calculate the S x K k-mer frequency matrix, M, where for subject *i, and k-mer j,* M_{ij} is the number of times k-mer *j* appears in $\boldsymbol{x_i}$, scaled by the total number of k-mers in $\boldsymbol{x_i}$:
$$M_{ij} = \frac{\sum_l I_{x_{il}=k_j}}{\sum_l \sum_m I_{x_{il}=k_m}}$$
and I is the indicator function.
Perform dimensionality reduction on M
Use Principal Component Analysis (PCA) to obtain the p principal components which best capture the differences between cases and controls.
If $\boldsymbol{w_i} = w_{i,1}, \dots, w_{i,K}$ is the i^{th} eigenvector of the covariance matrix of M, then for sample *s,* the *i*th principal component is given by $M_s \cdot \boldsymbol{w_i}$, where M_s is the s^{th} row of *M.*
Use contrastive PCA (cPCA) to transform the data onto a plane which captures the highest amount of variance in the cases compared to the controls.
Other methods of dimensionality reduction, such as factor analysis or variable selection
Train classification algorithm on the dimensionality reduced data set and use learnt parameters to determine disease status of new samples
Use supervised hierarchical clustering using algorithms such as hdbscan or agglomerative clustering with Ward's minimum variance, to learn the parameters which best distinguish cases from controls and can be used to classify new samples.
Discriminant analysis (including linear and quadratic discriminant analysis) to identify the features which separate the cases and controls. These features can then be applied to new samples.
Other methods of classification such as nearest neighbor algorithms and neural networks.

Figure 2 b

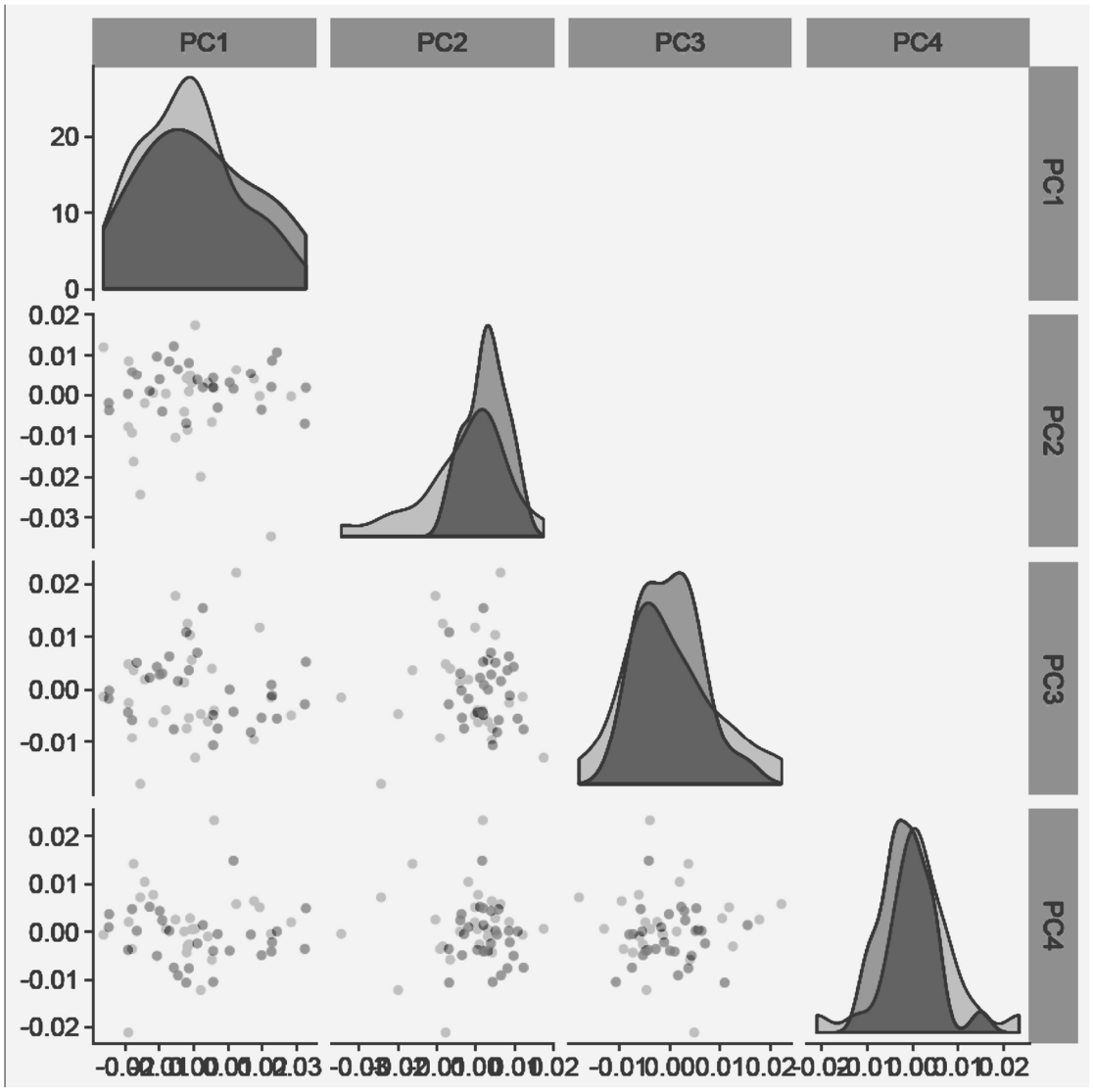

Figure 2c

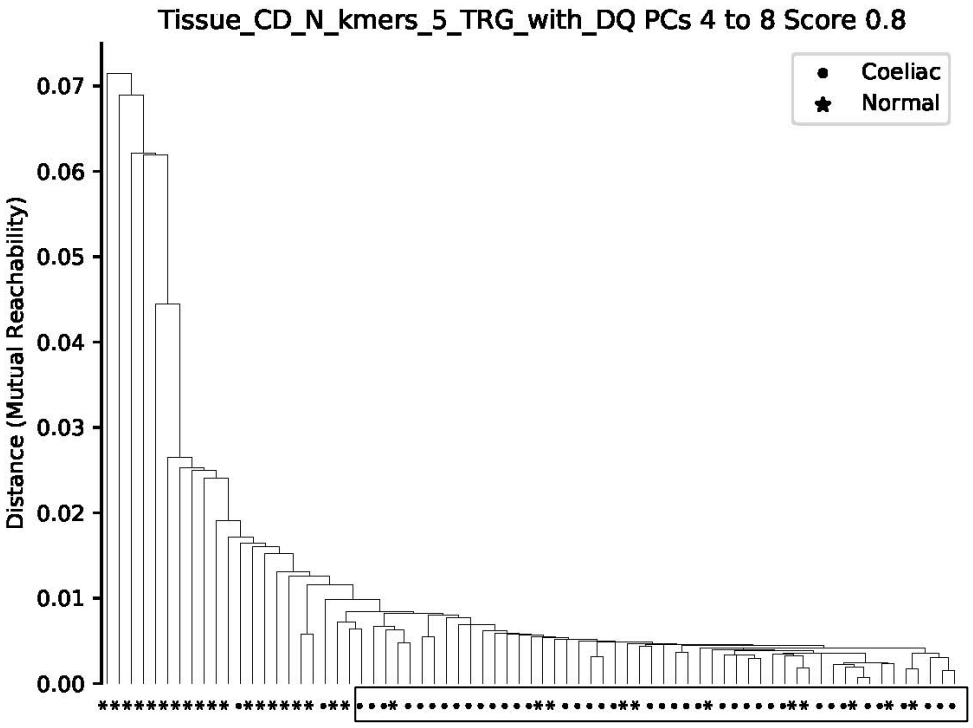

Figure 3a

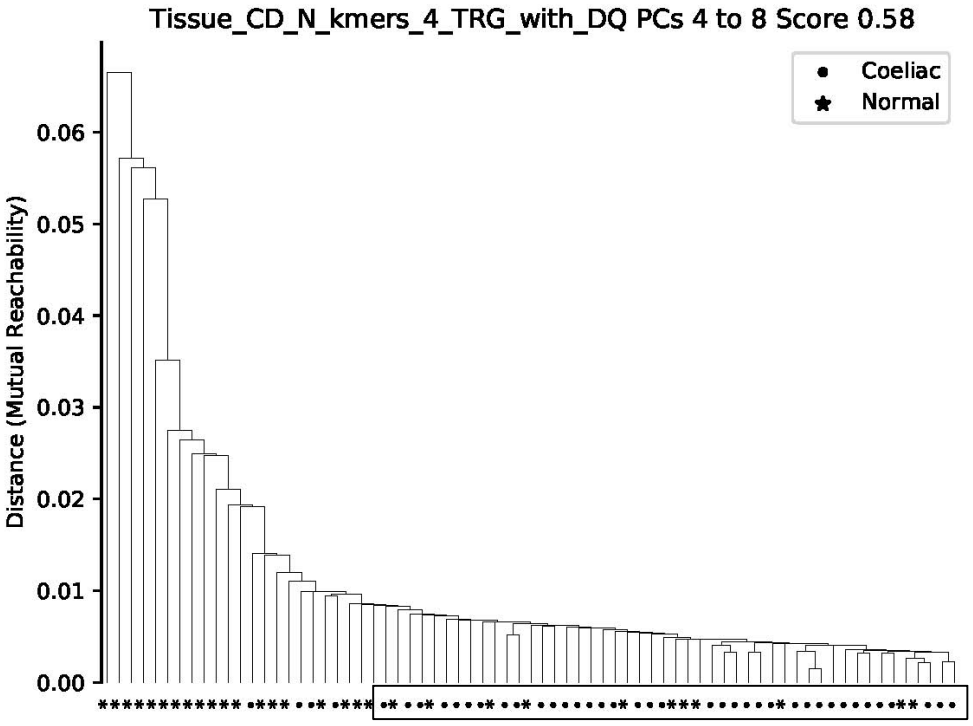

Figure 3b

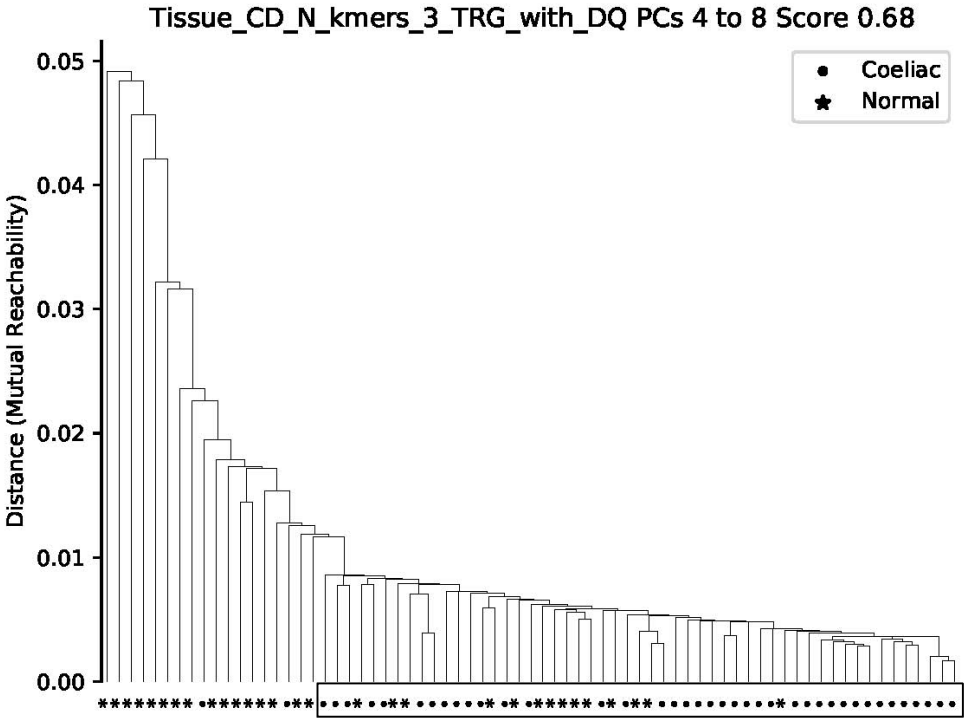

Figure 3c

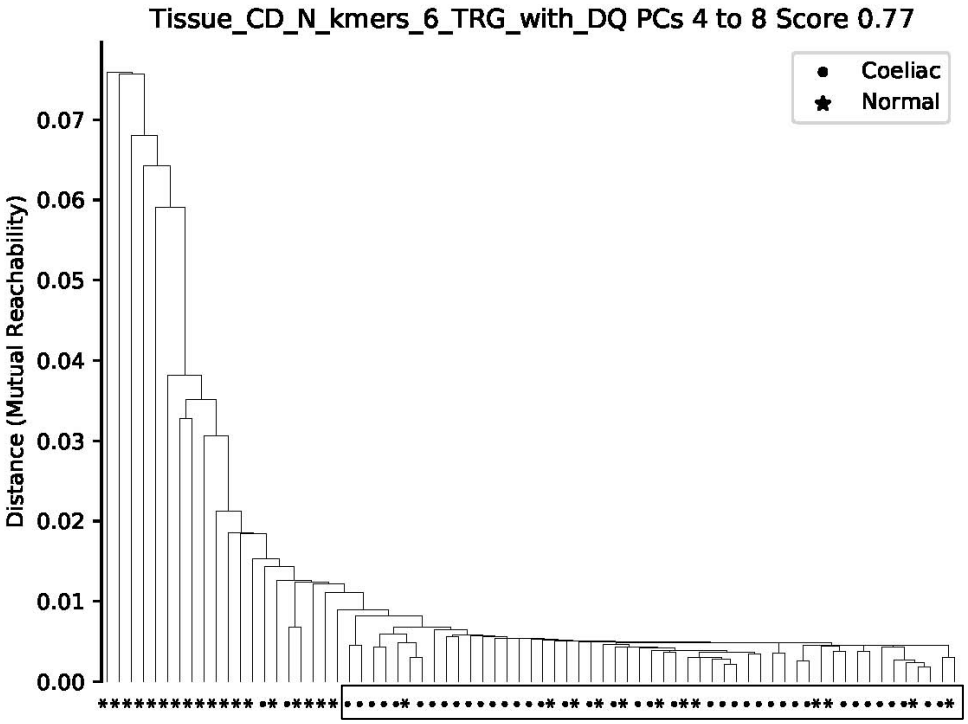

Figure 3d

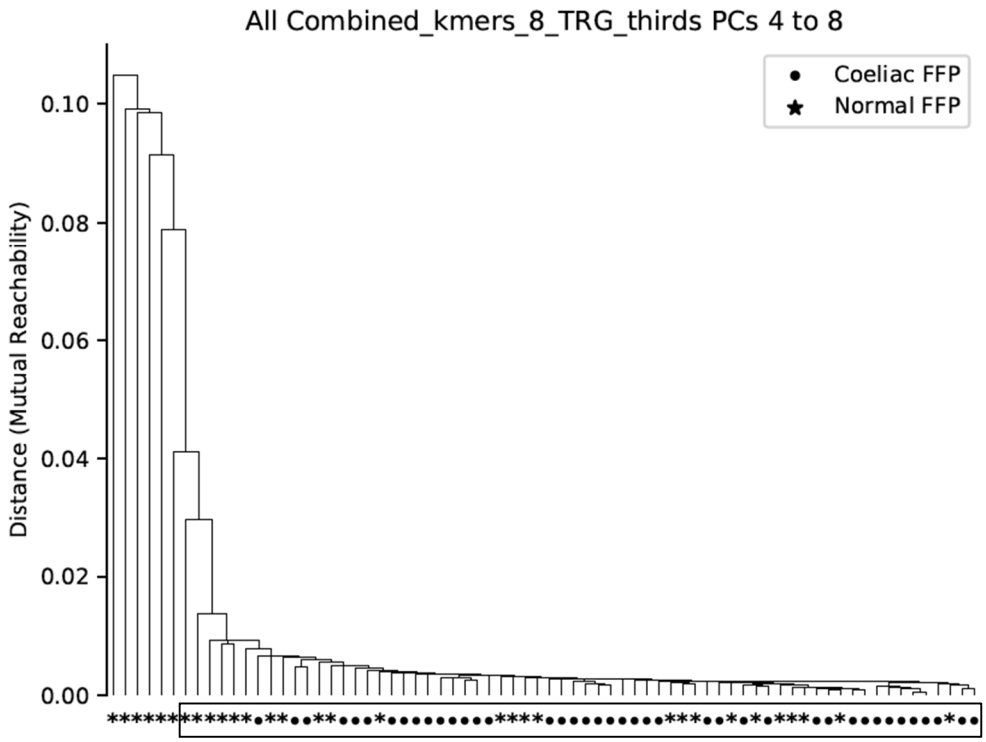

Figure 3e

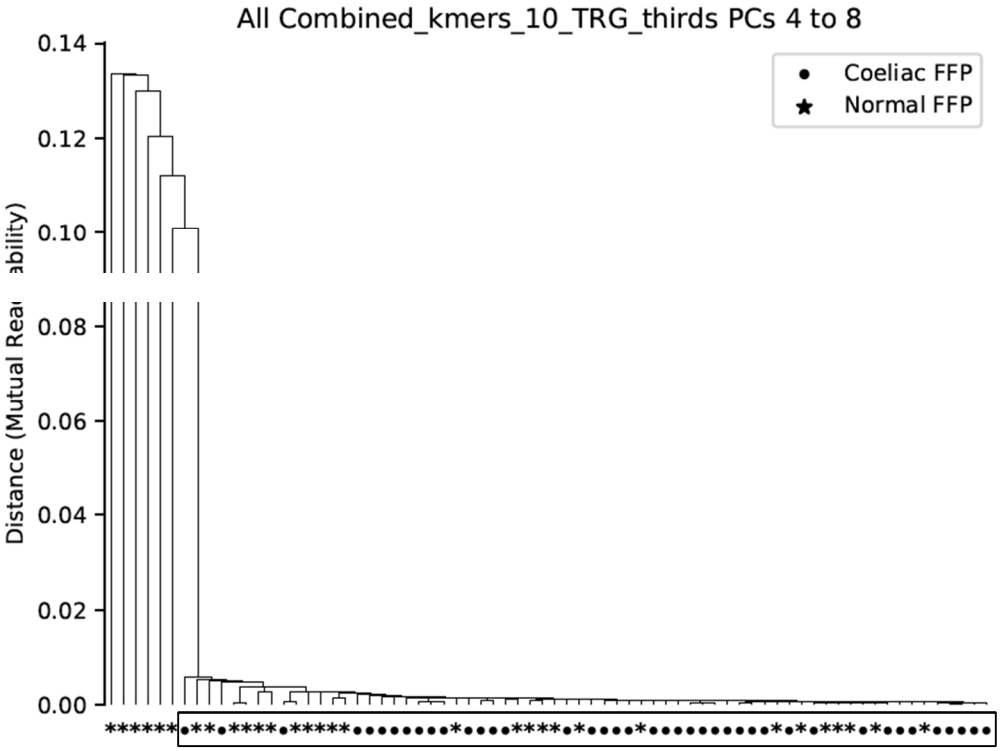

Figure 3f

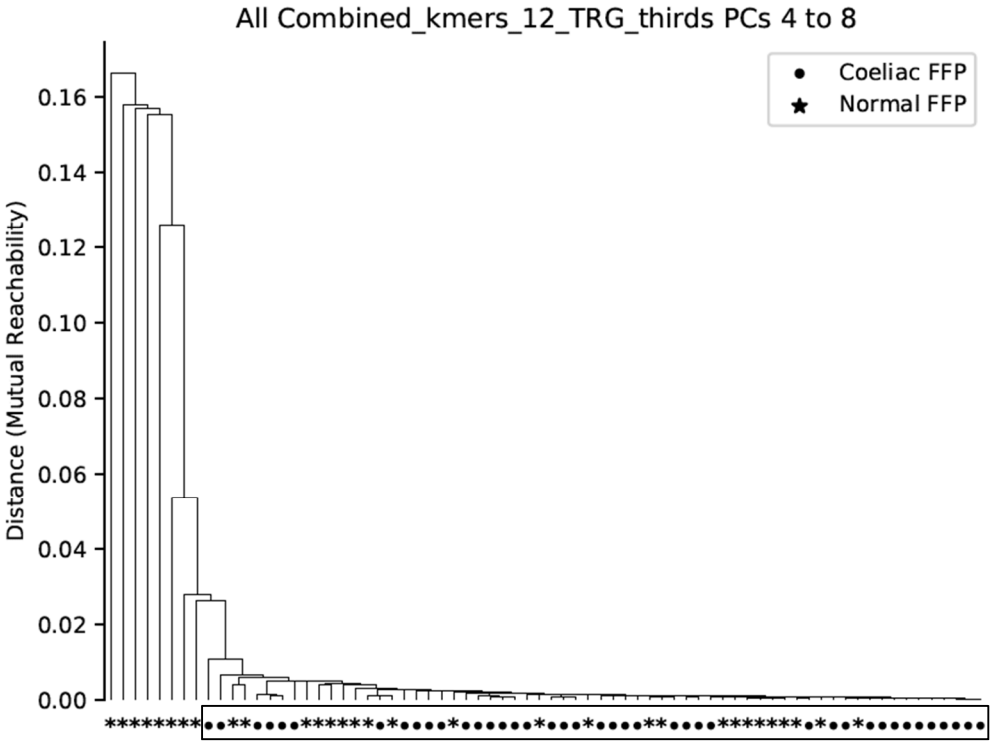
All Combined_kmers_12_TRG_thirds PCs 4 to 8
Coeliac FFP
Normal FFP
Distance (Mutual Reachability)
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00

Figure 3g

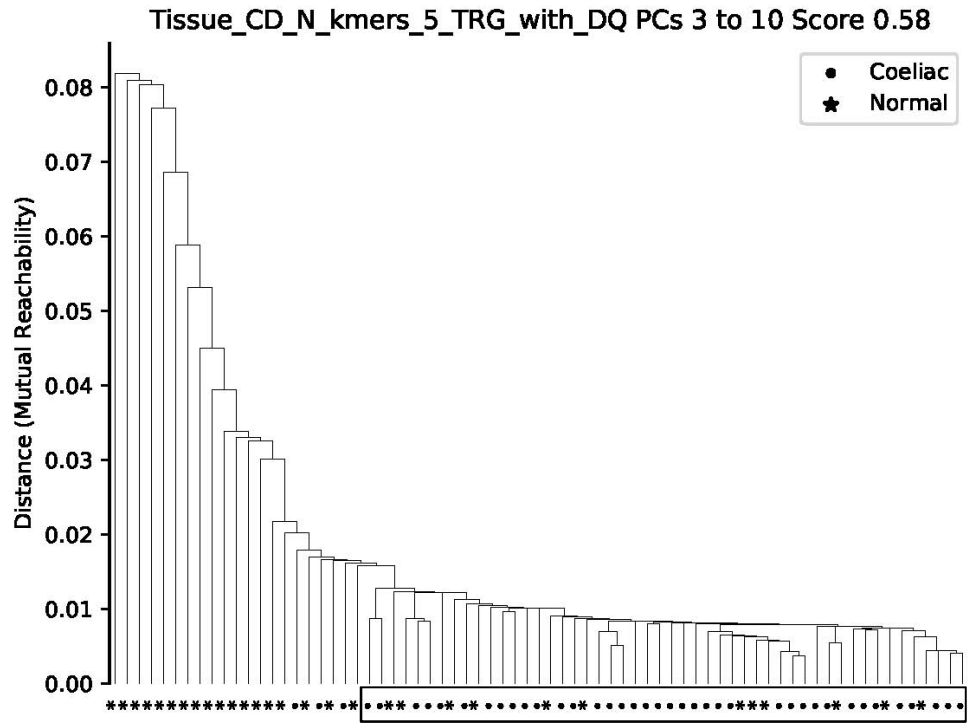

Figure 4a

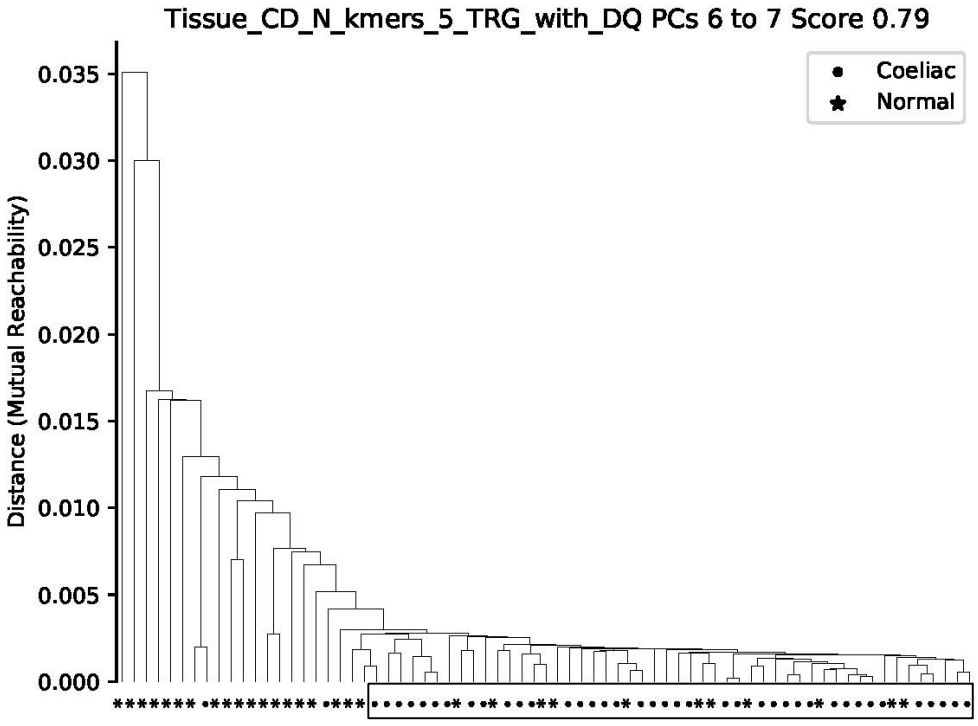

Figure 4b

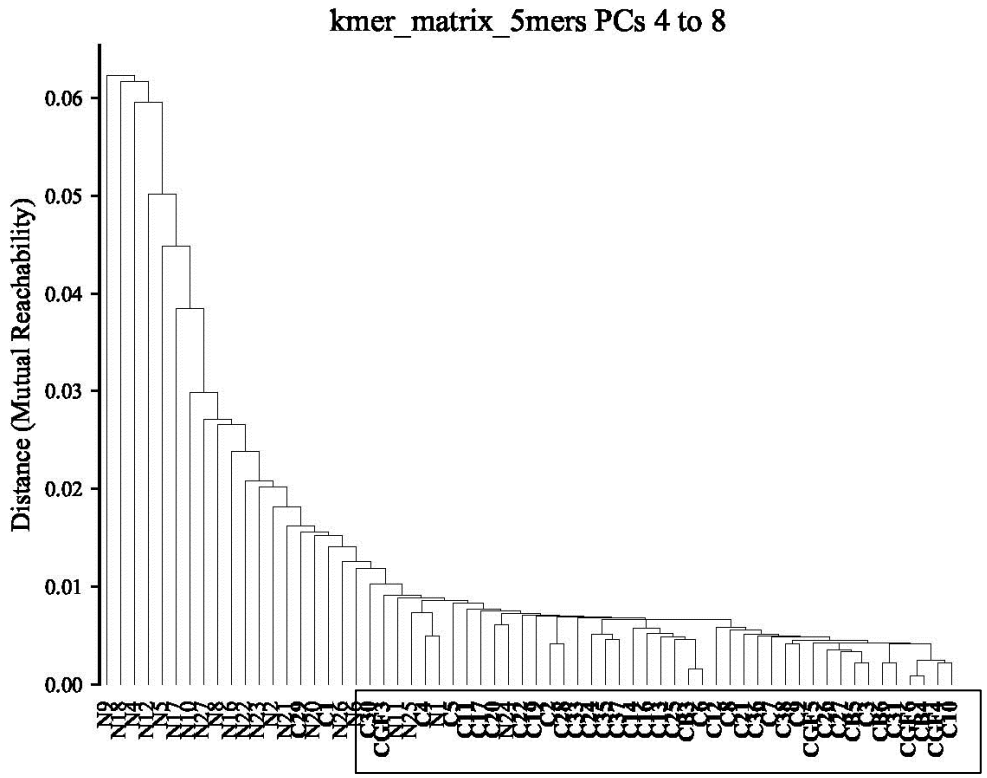

Figure 5

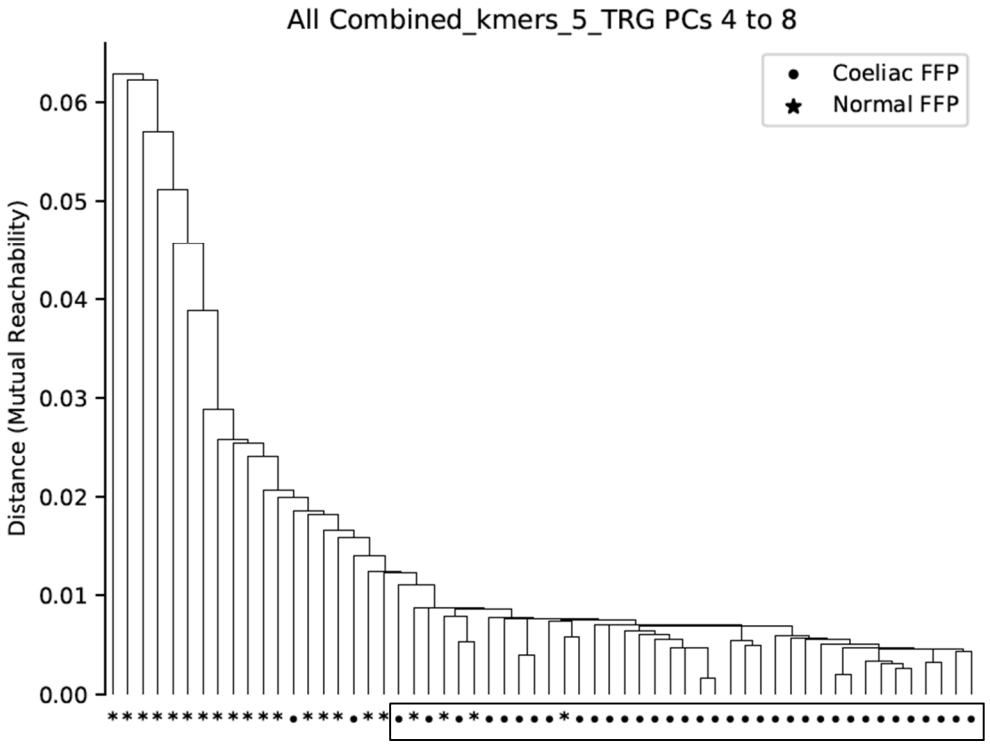

Figure 6a

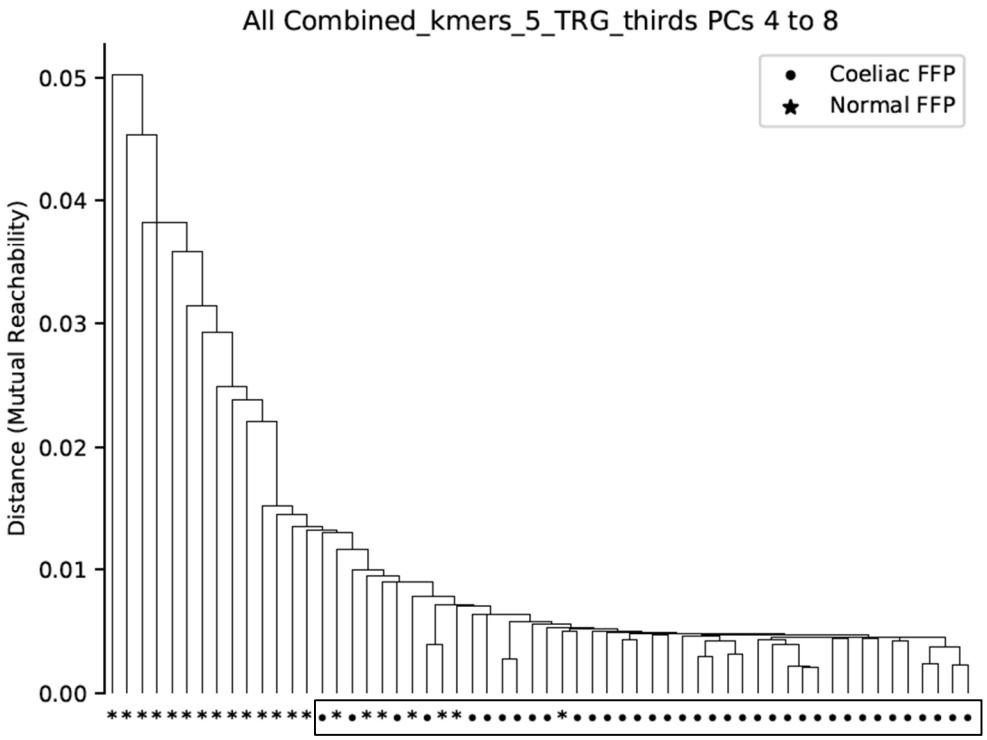
All Combined_kmers_5_TRG_thirds PCs 4 to 8
Coeliac FFP
Normal FFP
Distance (Mutual Reachability)
0.05
0.04
0.03
0.02
0.01
0.00

Figure 6b

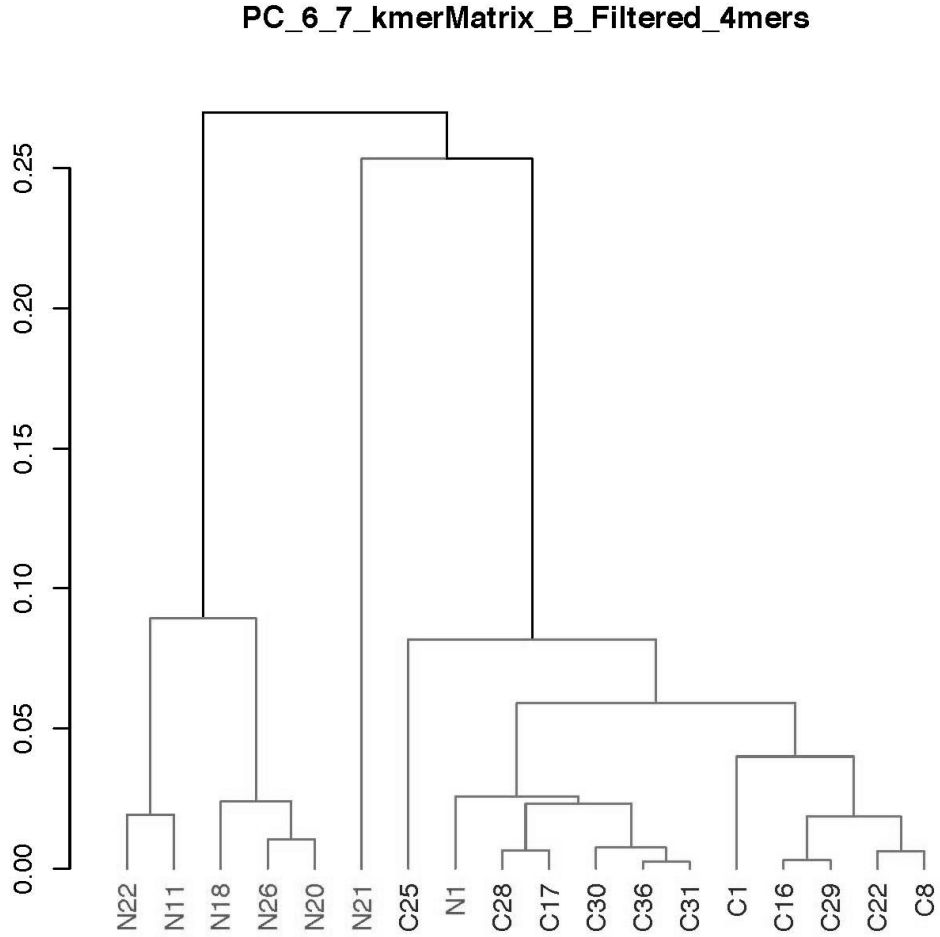
PC_6_7_kmerMatrix_B_Filtered_4mers
0.25
0.20
0.15
0.10
0.05
0.00
N22
N11
N18
N26
N20
N21
C25
N1
C28
C17
C30
C36
C31
C1
C16
C29
C22
C8

Figure 7a

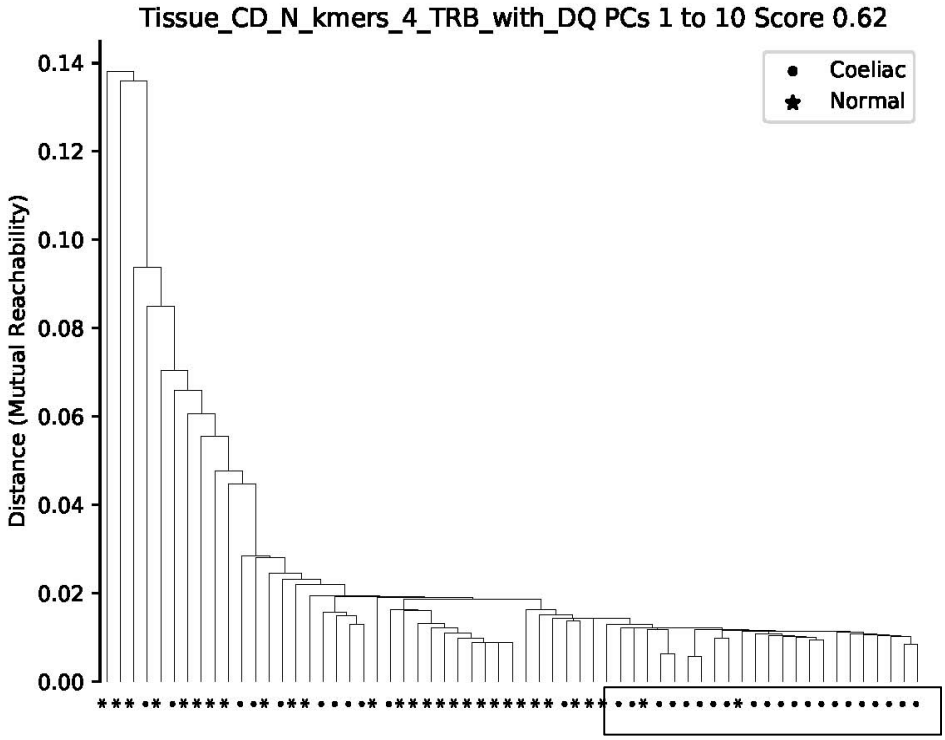

Figure 7b

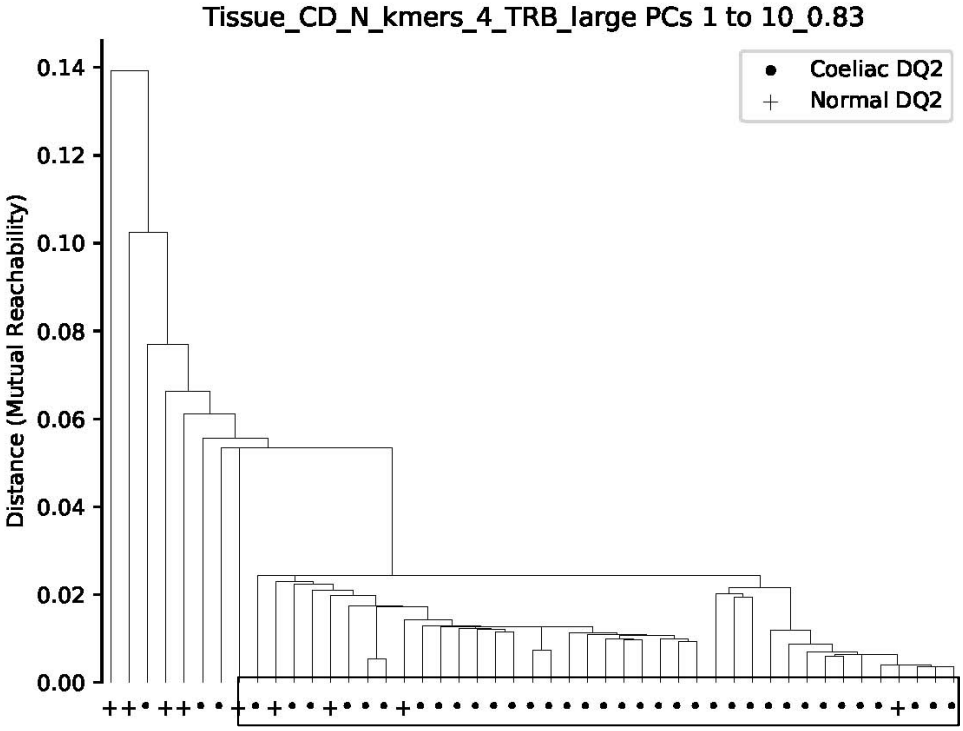

Figure 7c

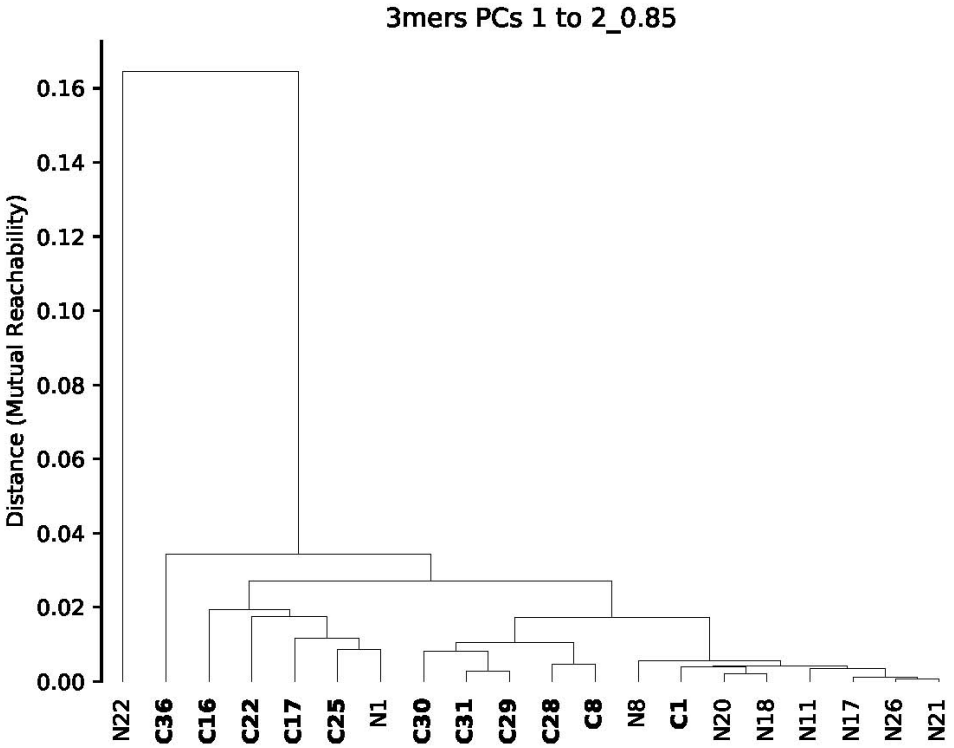

Figure 7d

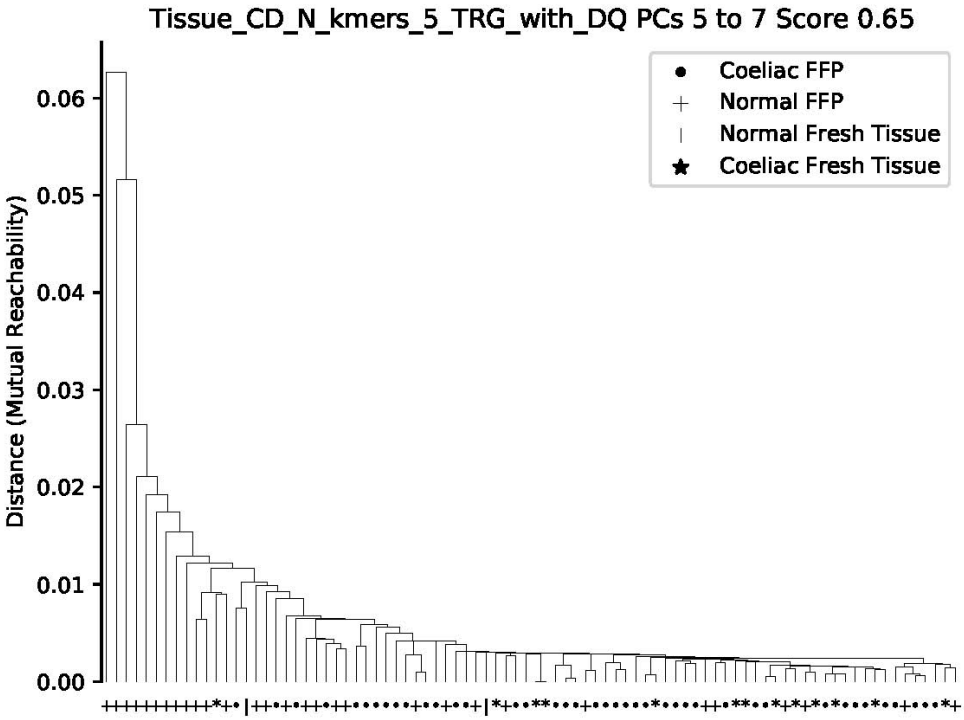

Figure 8

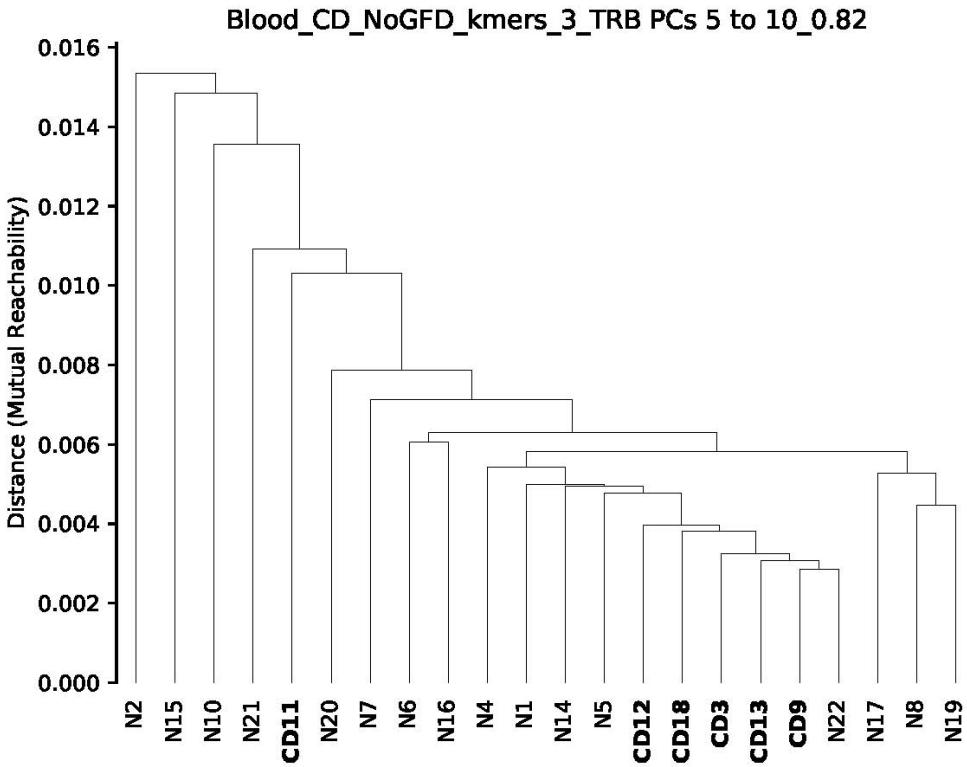

Figure 9

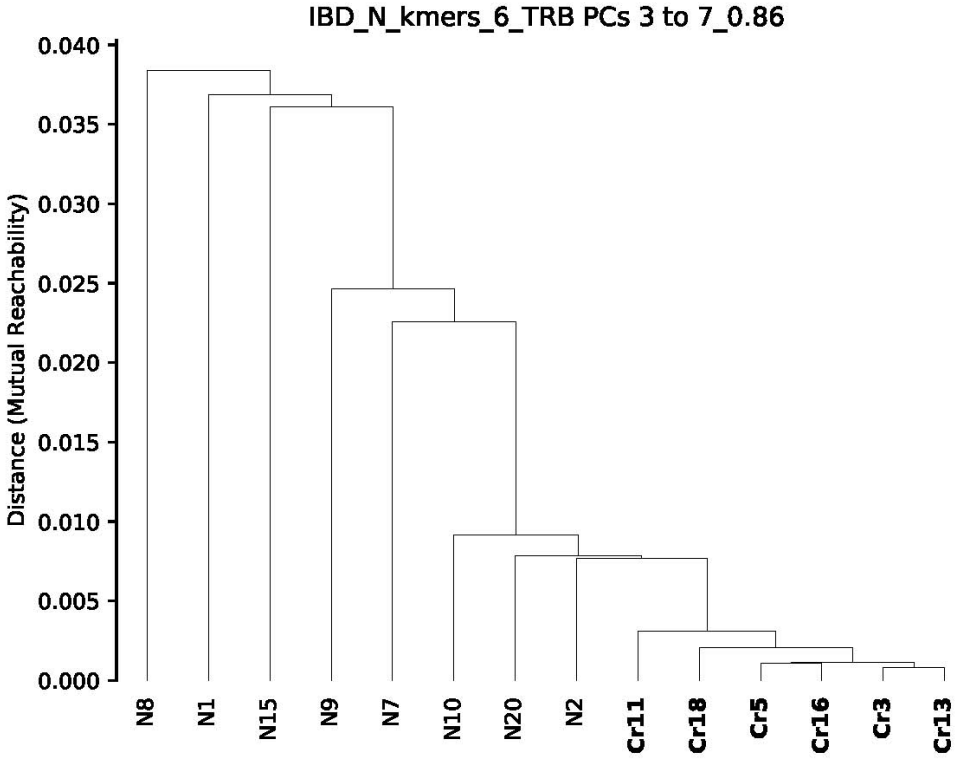

Figure 10

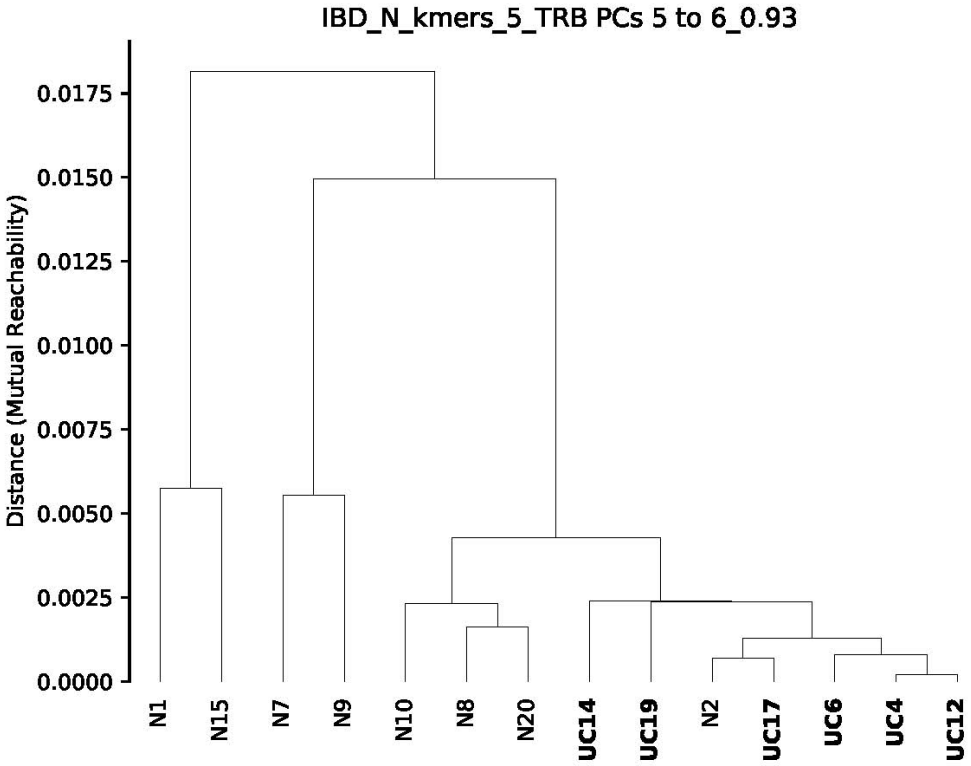

Figure 11a.

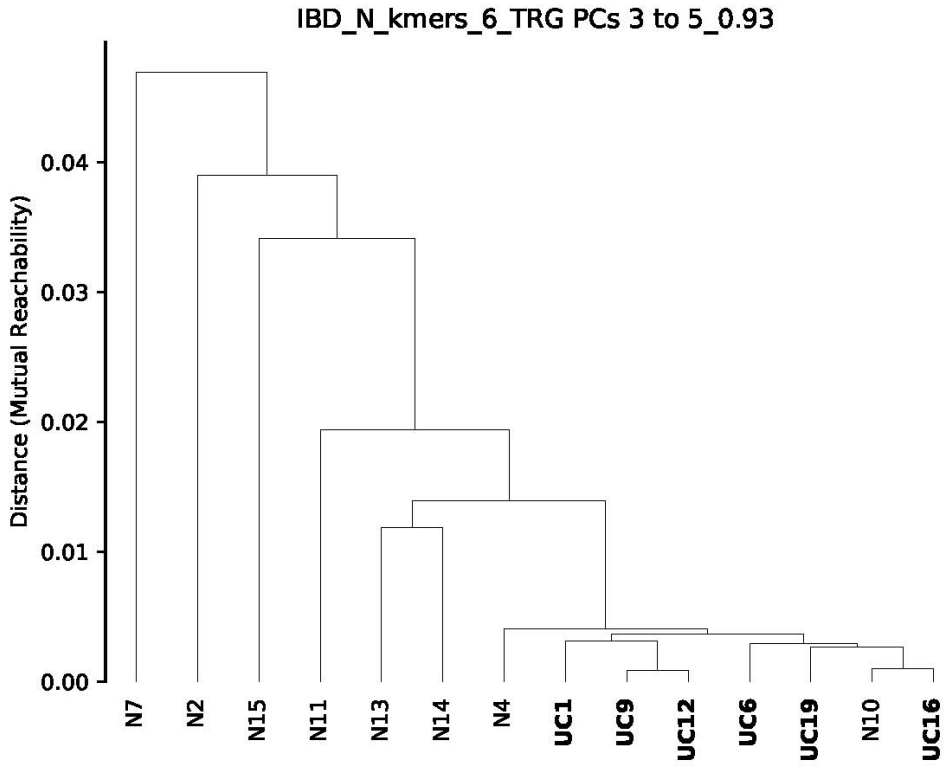

Figure 11b

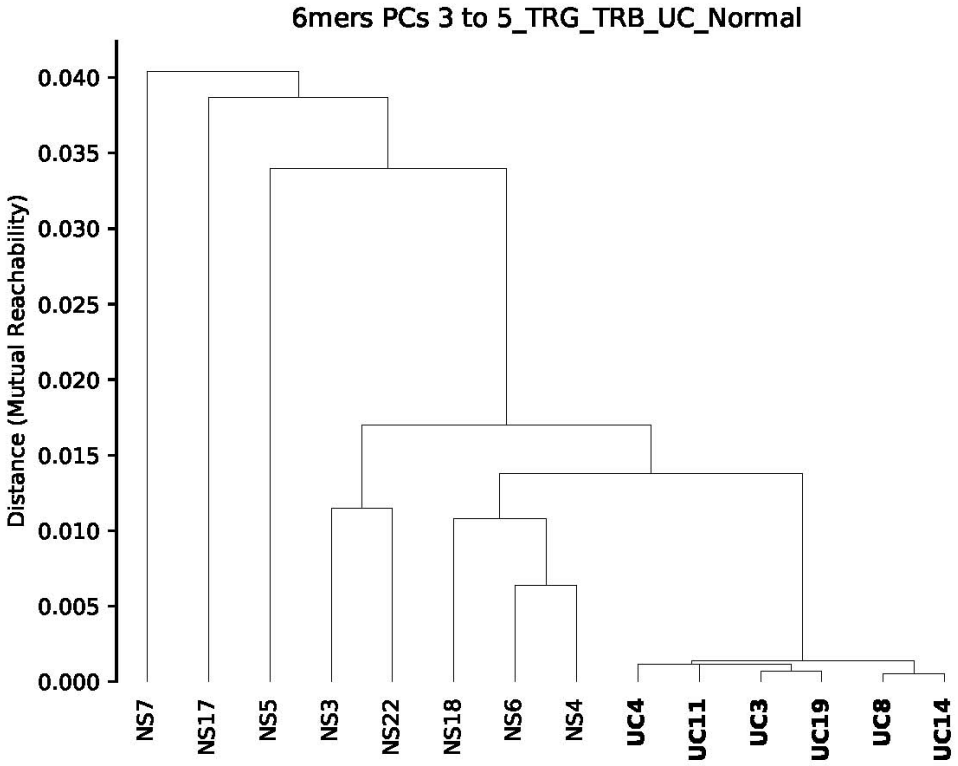

Figure 11c

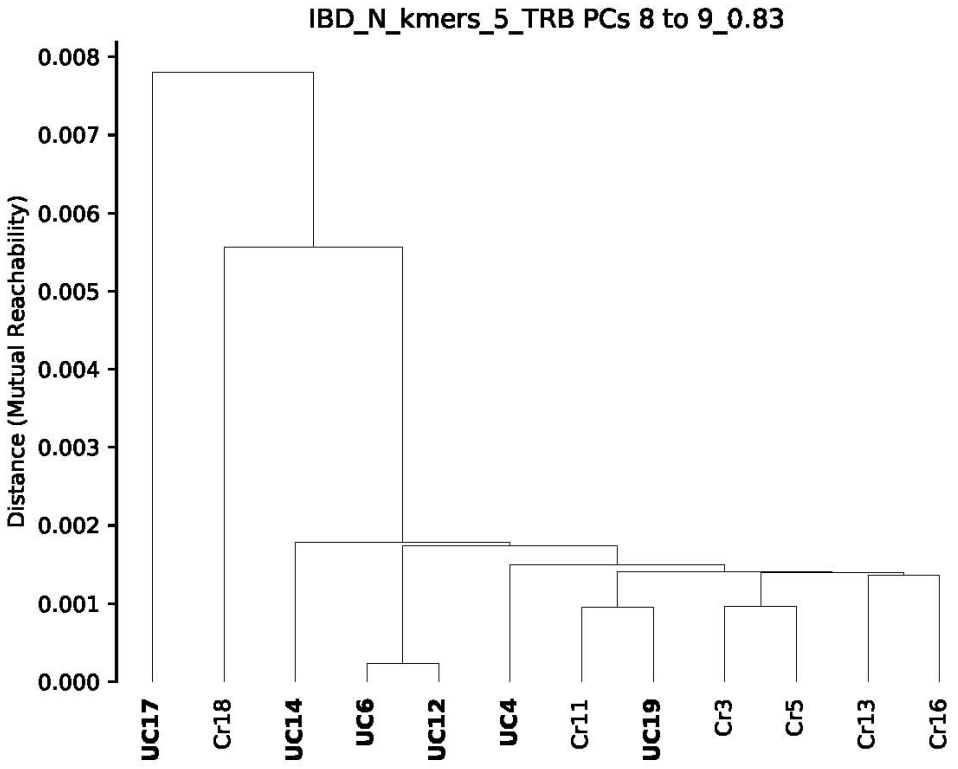

Figure 12

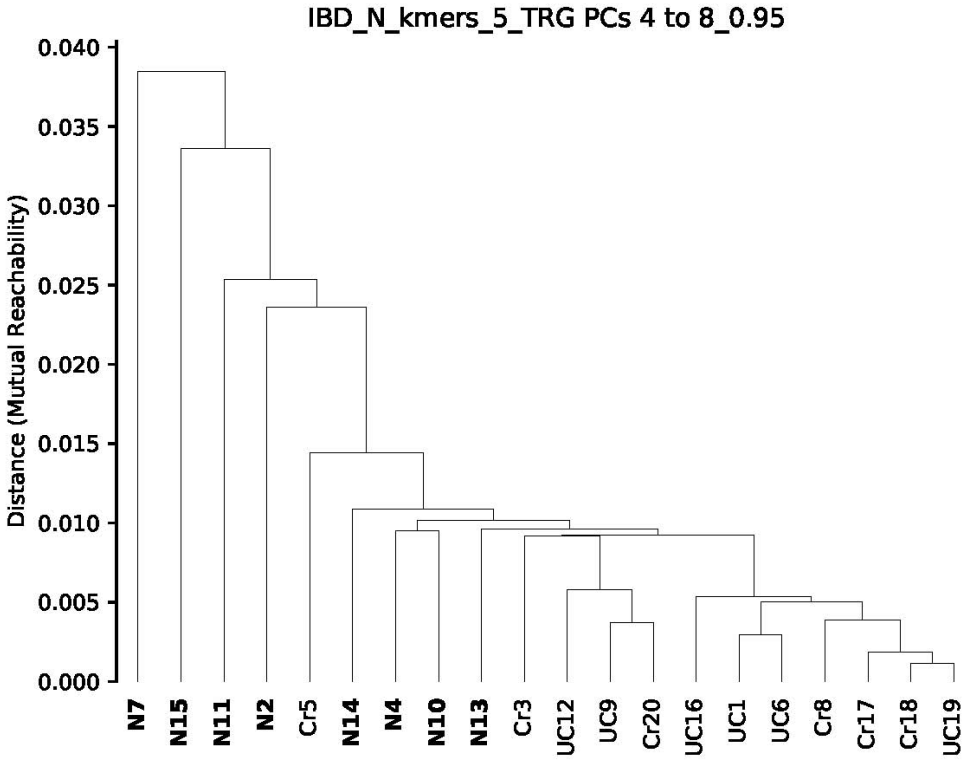

Figure 13

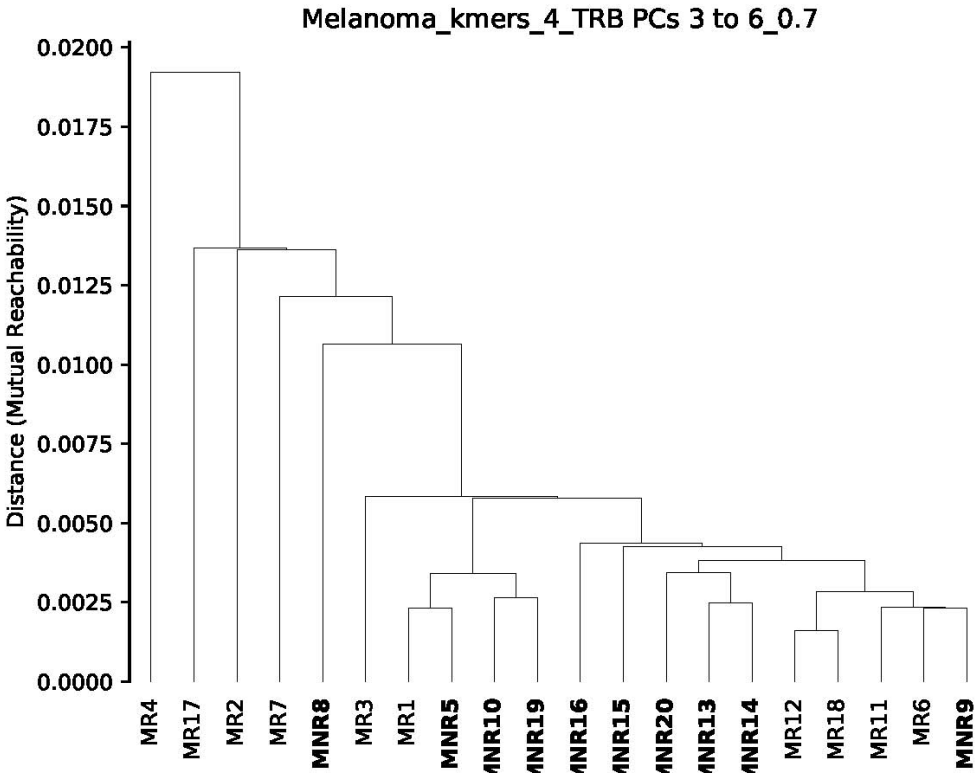

Figure 14

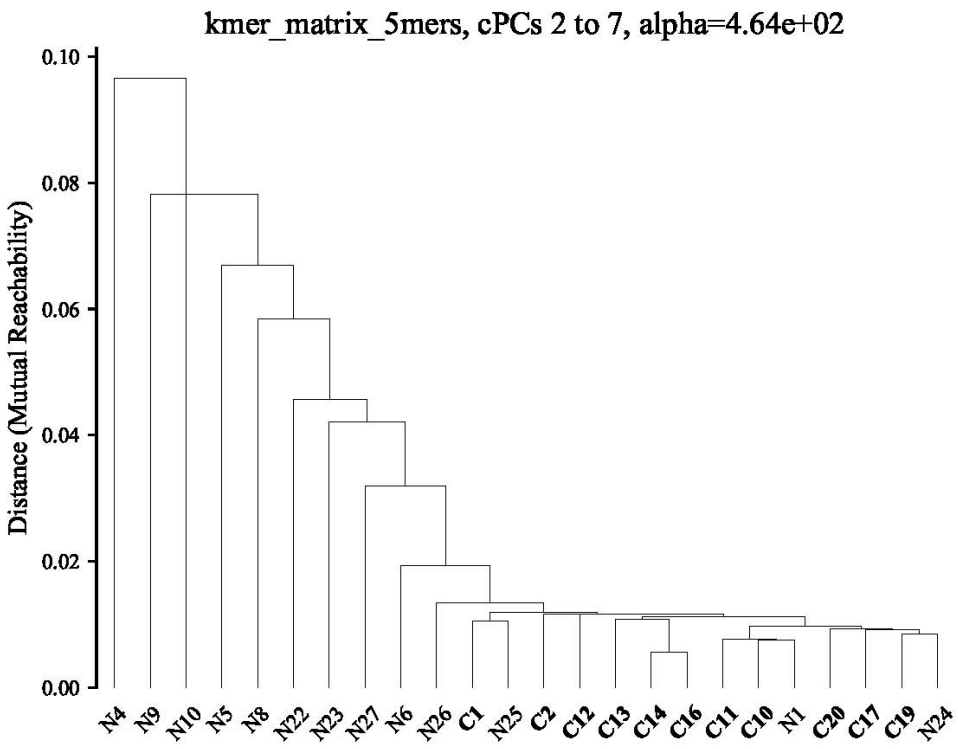
kmer_matrix_5mers, cPCs 2 to 7, alpha=4.64e+02
Distance (Mutual Reachability)
0.10
0.08
0.06
0.04
0.02
0.00
N4
N9
N10
N5
N8
N22
N23
N27
N6
N26
C1
N25
C2
C12
C13
C14
C16
C11
C10
N1
C20
C17
C19
N24

Figure 15.

Principal components	Score (k-mers without positional annotation)	New sample prediction accuracy (without positional annotation)	Overall Accuracy (k-mers without positional annotation)	Score (k-mers with positional annotation)	New sample prediction accuracy (with positional annotation)	Overall Accuracy (k-mers with positional annotation)
1 to 2	0.778	1.00	0.806	0.815	1.00	0.839
1 to 3	0.833	1.00	0.855	0.833	1.00	0.855
1 to 4	0.870	1.00	0.887	0.833	1.00	0.855
1 to 5	0.815	1.00	0.839	0.852	1.00	0.871
1 to 6	0.889	1.00	0.903	0.889	1.00	0.903
1 to 7	0.926	1.00	0.935	0.889	1.00	0.903
1 to 8	0.926	1.00	0.935	0.907	1.00	0.919
1 to 9	0.889	1.00	0.903	0.907	1.00	0.919
1 to 10	0.907	1.00	0.919	0.907	1.00	0.919
2 to 3	0.759	1.00	0.790	0.796	1.00	0.823
2 to 4	0.796	1.00	0.823	0.833	1.00	0.855
2 to 5	0.796	1.00	0.823	0.852	1.00	0.871
2 to 6	0.852	1.00	0.871	0.870	1.00	0.887
2 to 7	0.889	1.00	0.903	0.870	1.00	0.887
2 to 8	0.926	1.00	0.935	0.907	1.00	0.919
2 to 9	0.926	1.00	0.935	0.907	1.00	0.919
2 to 10	0.926	1.00	0.935	0.907	1.00	0.919
3 to 4	0.759	1.00	0.790	0.759	1.00	0.790
3 to 5	0.815	1.00	0.839	0.852	1.00	0.871
3 to 6	0.852	1.00	0.871	0.833	1.00	0.855
3 to 7	0.889	1.00	0.903	0.852	1.00	0.871
3 to 8	0.926	1.00	0.935	0.926	1.00	0.935
3 to 9	0.926	1.00	0.935	0.907	1.00	0.919
3 to 10	0.944	1.00	0.952	0.907	1.00	0.919
4 to 5	0.796	1.00	0.823	0.852	1.00	0.871
4 to 6	0.870	1.00	0.887	0.852	1.00	0.871
4 to 7	0.870	1.00	0.887	0.852	1.00	0.871
4 to 8	0.944	1.00	0.952	0.926	1.00	0.935
4 to 9	0.926	1.00	0.935	0.926	1.00	0.935
4 to 10	0.907	1.00	0.919	0.907	1.00	0.919
5 to 6	0.796	1.00	0.823	0.815	1.00	0.839
5 to 7	0.833	1.00	0.855	0.815	1.00	0.839
5 to 8	0.907	1.00	0.919	0.907	1.00	0.919
5 to 9	0.907	1.00	0.919	0.907	1.00	0.919
5 to 10	0.907	1.00	0.919	0.926	1.00	0.935
6 to 7	0.833	1.00	0.855	0.852	1.00	0.871
6 to 8	0.907	1.00	0.919	0.907	1.00	0.919
6 to 9	0.907	1.00	0.919	0.907	1.00	0.919
6 to 10	0.889	1.00	0.903	0.889	1.00	0.903
7 to 8	0.833	1.00	0.855	0.852	1.00	0.871
7 to 9	0.852	1.00	0.871	0.889	1.00	0.903
7 to 10	0.852	1.00	0.871	0.852	1.00	0.871
8 to 9	0.815	1.00	0.839	0.852	1.00	0.871
8 to 10	0.852	1.00	0.871	0.833	1.00	0.855
9 to 10	0.796	1.00	0.823	0.796	1.00	0.823

Figure 16.

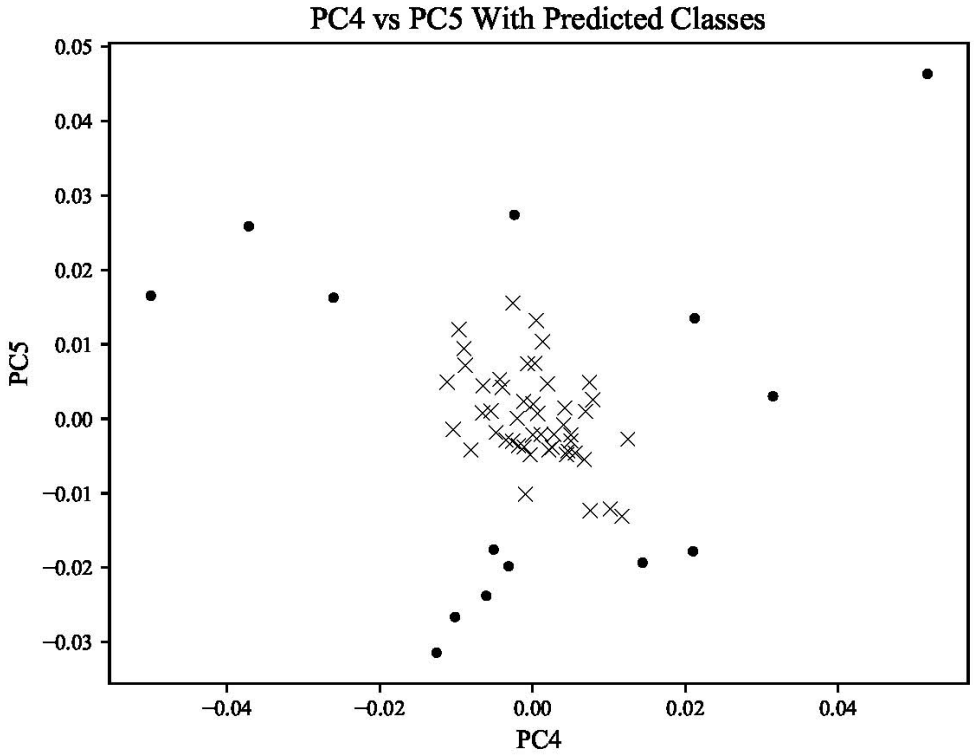

Figure 17a.

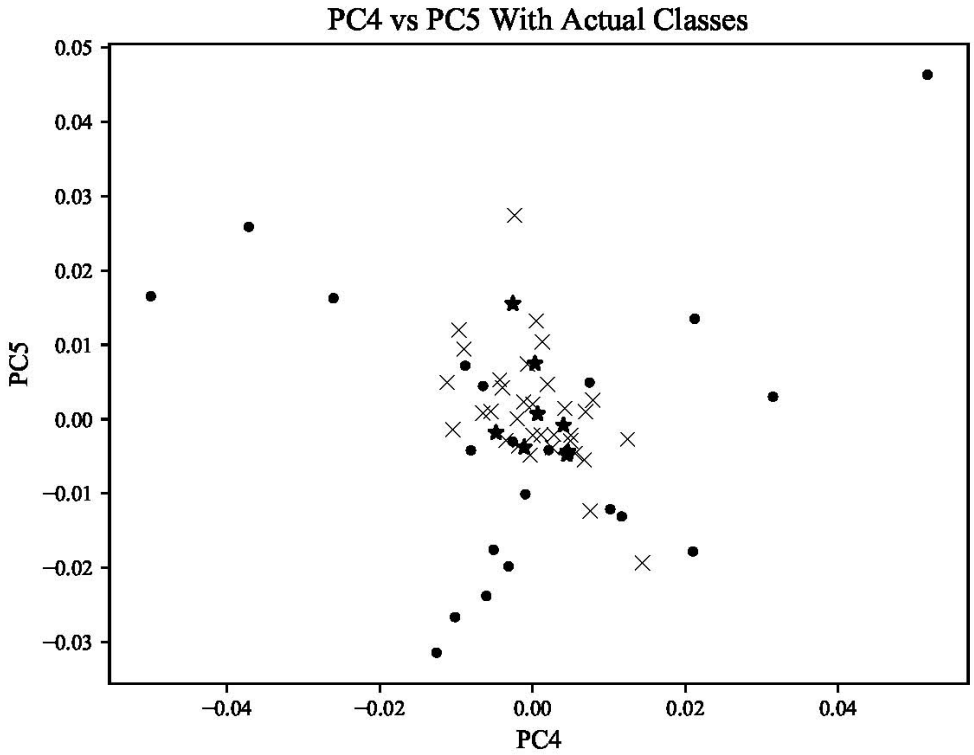

Figure 17b.

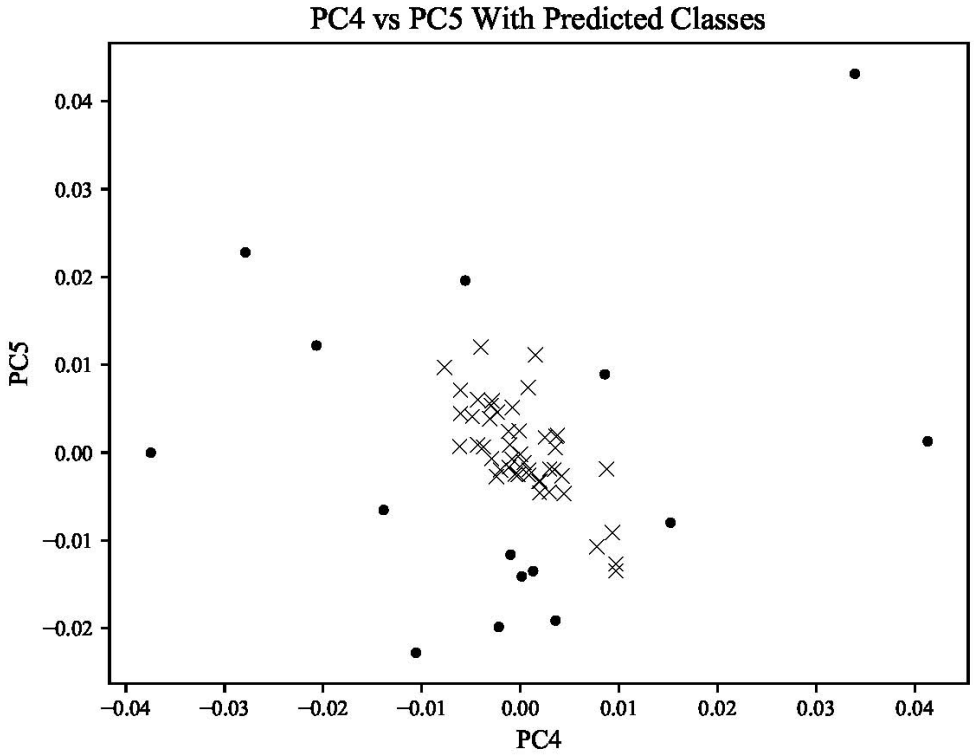

Figure 17c.

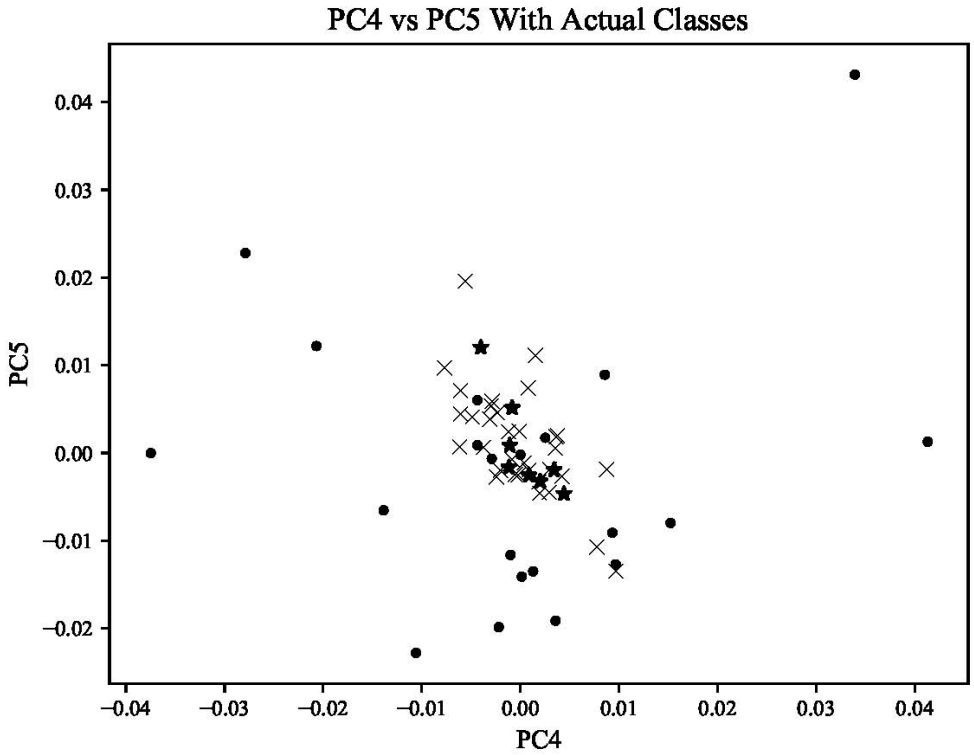

Figure 17d.

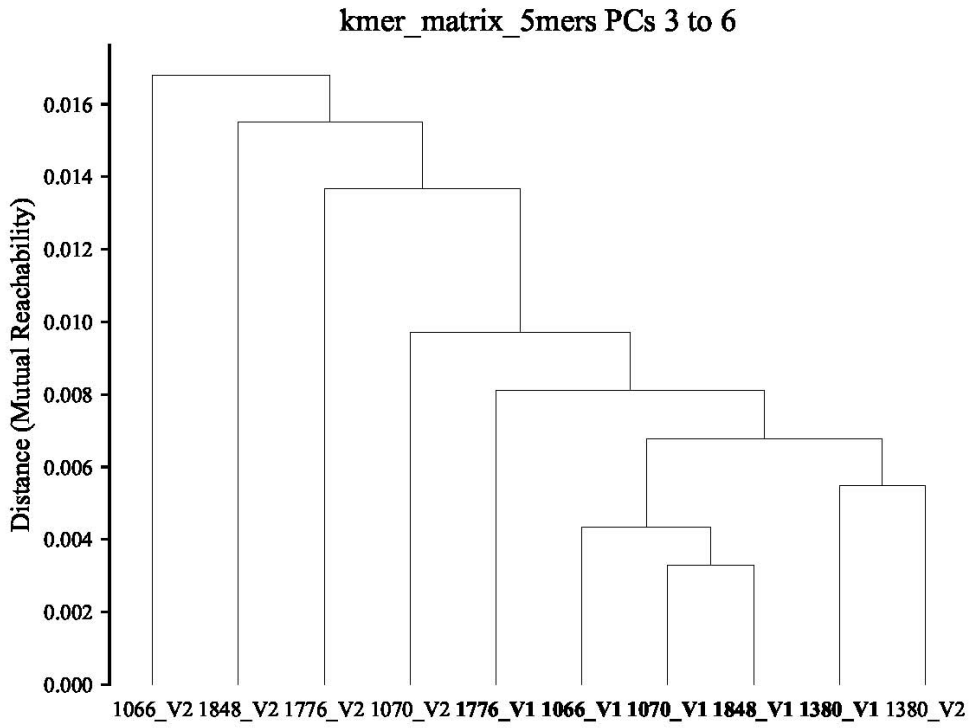
kmer_matrix_5mers PCs 3 to 6
Distance (Mutual Reachability)
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
1066_V2
1848_V2
1776_V2
1070_V2
1776_V1
1066_V1
1070_V1
1848_V1
1380_V1
1380_V2

Figure 18