I2A2 - Genomic Analysis Challenge

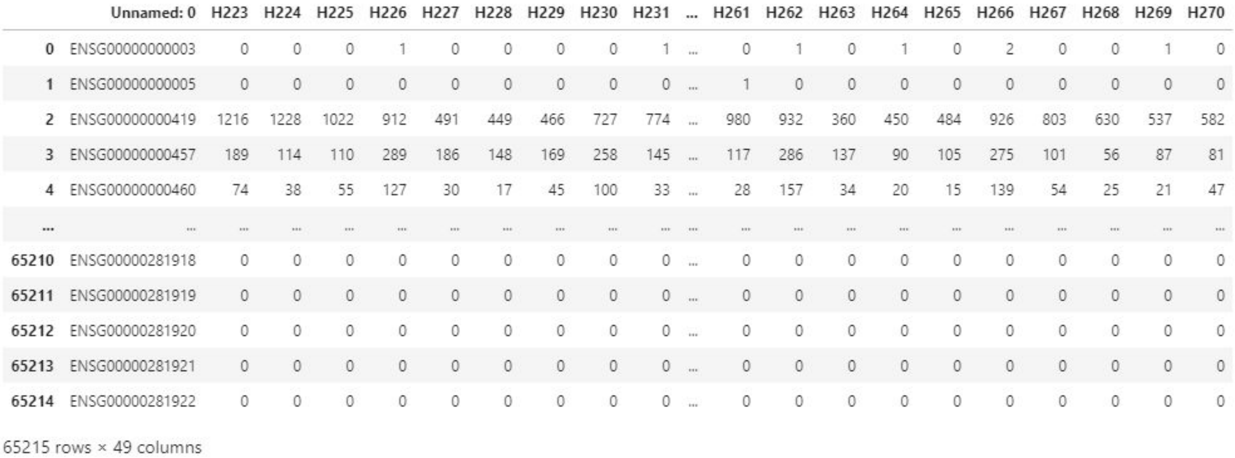
Student: Weld Lucas Cunha

Problem definition

Identify which people are relatives in a database composed of the genomic data of 48 people.

Data exploration

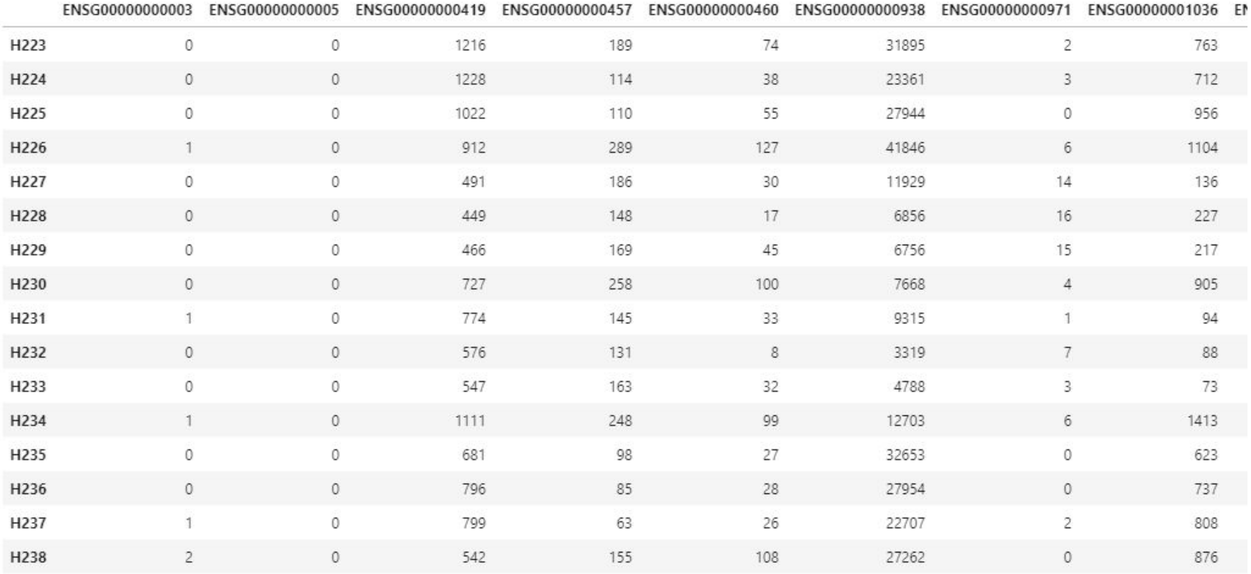
The database is composed of 65215 different genes from 48 people.



	Unnamed: 0	H223	H224	H225	H226	H227	H228	H229	H230	H231	...	H261	H262	H263	H264	H265	H266	H267	H268	H269	H270
0	ENSG00000000003	0	0	0	1	0	0	0	0	1	...	0	1	0	1	0	2	0	0	1	0
1	ENSG00000000005	0	0	0	0	0	0	0	0	0	...	1	0	0	0	0	0	0	0	0	0
2	ENSG00000000419	1216	1228	1022	912	491	449	466	727	774	...	980	932	360	450	484	926	803	630	537	582
3	ENSG00000000457	189	114	110	289	186	148	169	258	145	...	117	286	137	90	105	275	101	56	87	81
4	ENSG00000000460	74	38	55	127	30	17	45	100	33	...	28	157	34	20	15	139	54	25	21	47
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
65210	ENSG00000281918	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
65211	ENSG00000281919	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
65212	ENSG00000281920	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
65213	ENSG00000281921	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
65214	ENSG00000281922	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

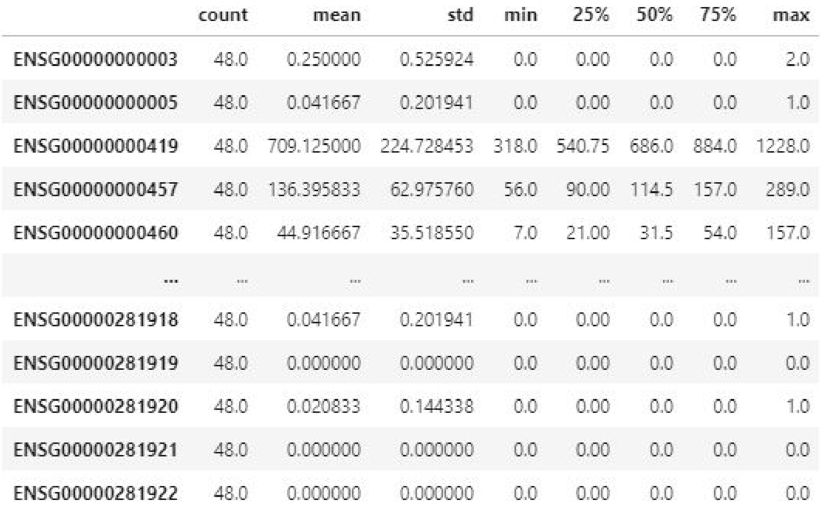
65215 rows × 49 columns

Transposing the table, to put the features (genes values) in the columns and the subjects info in the rows:



	ENSG00000000003	ENSG00000000005	ENSG00000000419	ENSG00000000457	ENSG00000000460	ENSG00000000938	ENSG00000000971	ENSG00000001036	EI
H223	0	0	1216	189	74	31895	2	763	
H224	0	0	1228	114	38	23361	3	712	
H225	0	0	1022	110	55	27944	0	956	
H226	1	0	912	289	127	41846	6	1104	
H227	0	0	491	186	30	11929	14	136	
H228	0	0	449	148	17	6856	16	227	
H229	0	0	466	169	45	6756	15	217	
H230	0	0	727	258	100	7668	4	905	
H231	1	0	774	145	33	9315	1	94	
H232	0	0	576	131	8	3319	7	88	
H233	0	0	547	163	32	4788	3	73	
H234	1	0	1111	248	99	12703	6	1413	
H235	0	0	681	98	27	32653	0	623	
H236	0	0	796	85	28	27954	0	737	
H237	1	0	799	63	26	22707	2	808	
H238	2	0	542	155	108	27262	0	876	

A simple overview of the dataset shows us that the values of the genes vary a lot and have different ranges of values:



	count	mean	std	min	25%	50%	75%	max
ENSG00000000003	48.0	0.250000	0.525924	0.0	0.00	0.0	0.0	2.0
ENSG00000000005	48.0	0.041667	0.201941	0.0	0.00	0.0	0.0	1.0
ENSG00000000419	48.0	709.125000	224.728453	318.0	540.75	686.0	884.0	1228.0
ENSG00000000457	48.0	136.395833	62.975760	56.0	90.00	114.5	157.0	289.0
ENSG00000000460	48.0	44.916667	35.518550	7.0	21.00	31.5	54.0	157.0
...	...	...	...	...	...	...	...	...
ENSG00000281918	48.0	0.041667	0.201941	0.0	0.00	0.0	0.0	1.0
ENSG00000281919	48.0	0.000000	0.000000	0.0	0.00	0.0	0.0	0.0
ENSG00000281920	48.0	0.020833	0.144338	0.0	0.00	0.0	0.0	1.0
ENSG00000281921	48.0	0.000000	0.000000	0.0	0.00	0.0	0.0	0.0
ENSG00000281922	48.0	0.000000	0.000000	0.0	0.00	0.0	0.0	0.0

Strategy Definition

- Apply some dimensionality reduction, due to the great number of genes (features) available;
- Cluster the data points to group the ones with more similarities;
- It's expected that some subjects in the dataset won't have relatives, so in this case, it's desirable to let them out of any cluster. Clustering algorithms like DBScan or HDBScan are capable of letting some samples "unclustered";
- Defining the *eps* value for the DBScan algorithm might be tricky sometimes, especially when the data is unknown, like in this case. Therefore, HDBScan might be a better option, since its algorithm is able to find the more stable clusters in the data without setting many parameters;
- Also, different techniques can be used for dimensionality reduction, like PCA, ICA, UMAP, etc. And there is no guarantee that one specific technique is the single best one. But, it is expected that similar samples will be still similar in the new features space for most of the techniques applied.

Summary: Different dimensionality reduction techniques will be applied, the HDBScan will be applied over the reduced space for each technique. In the second stage, all results will be summarized like in an ensemble classifier, but in this case, an ensemble clustering model.

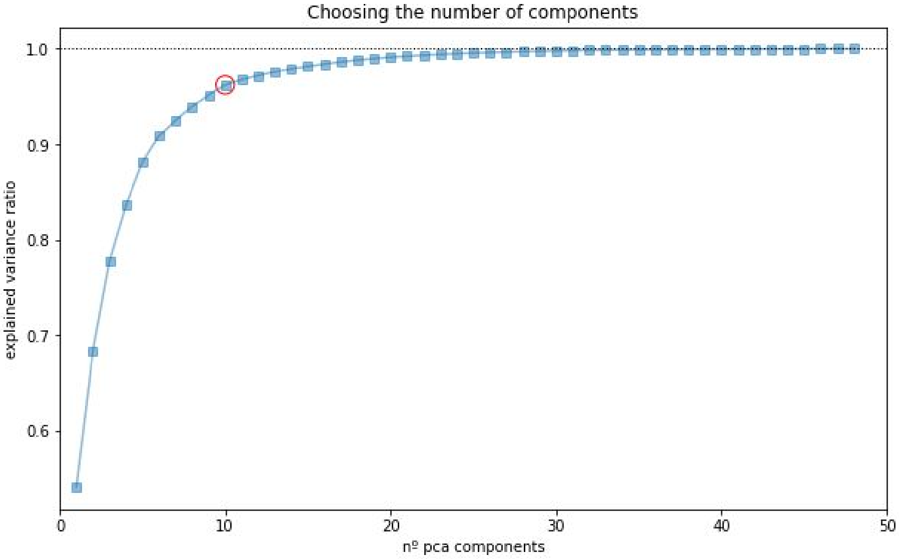
Implementation

The following dimensionality reduction techniques were chosen:

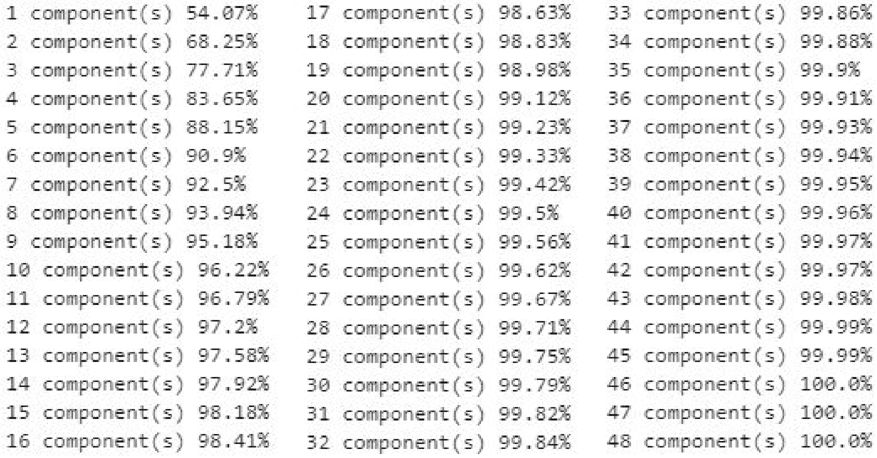
- Principal component analysis (PCA)
- Non-linear dimensionality reduction through Isometric Mapping (Isomap)
- Multidimensional scaling (MDS)
- Spectral embedding for non-linear dimensionality reduction (SE)
- Uniform Manifold Approximation and Projection (UMAP)

Setting the number of components

Our dataset has 48 rows and 65215 columns, therefore, PCA can produce up to 48 components in the new features space. However, in order to better calculate distances among the samples, we should choose as few components as possible, as long as we guarantee a good representation of the original data into the new features space. In order to determine the number of components, the PCA algorithm was chosen, and the following graph shows the cumulative explained variance ratio for different numbers of components chosen.



The explained variance ratio was used to choose the number of components that will be used in the clustering algorithm. According to the following explained variance ratio values, we chose 10 components, in this case, 96.22% of all variance is explained.

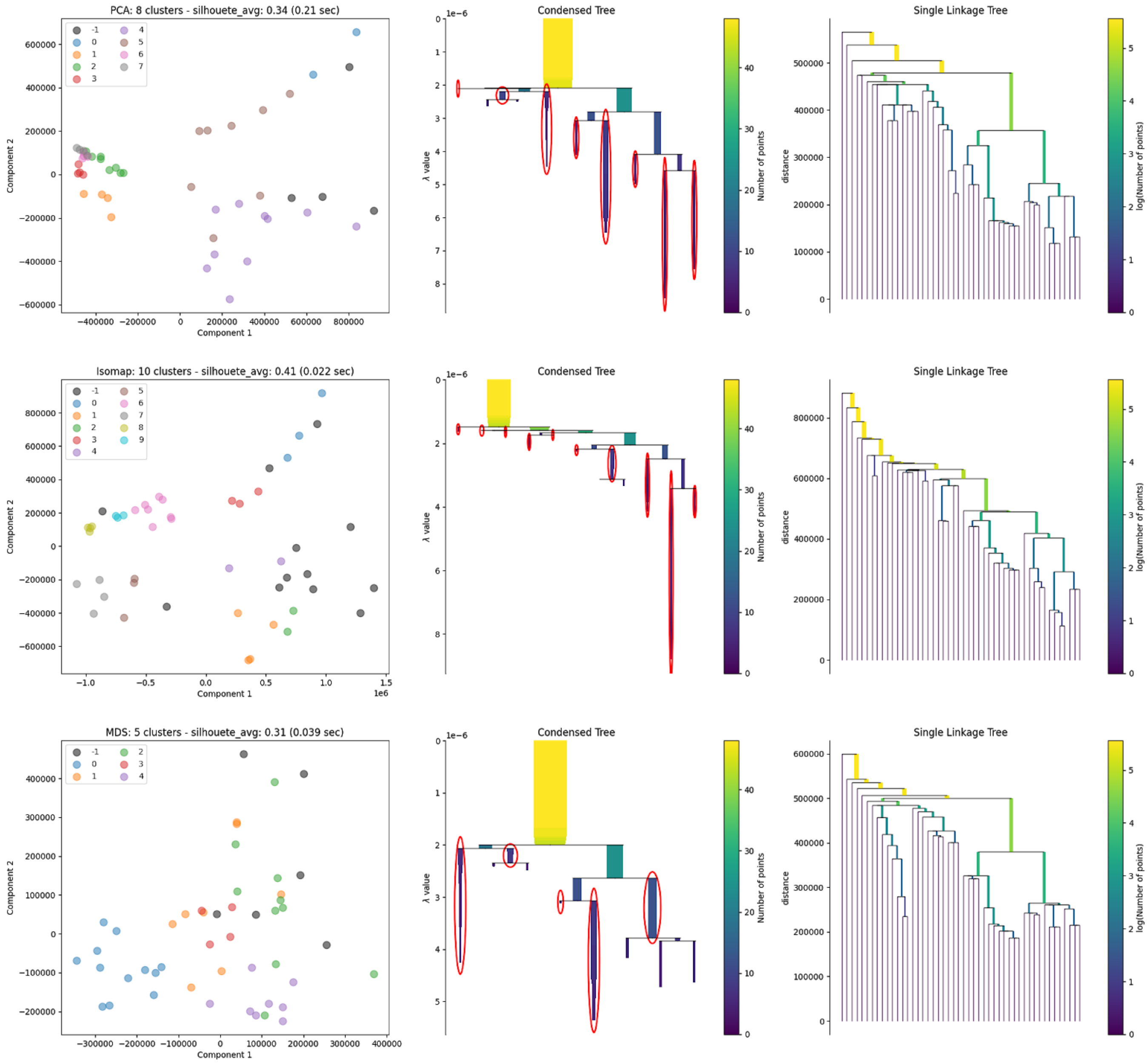

```
 1 component(s) 54.07%    17 component(s) 98.63%   33 component(s) 99.86%
 2 component(s) 68.25%    18 component(s) 98.83%   34 component(s) 99.88%
 3 component(s) 77.71%    19 component(s) 98.98%   35 component(s) 99.9%
 4 component(s) 83.65%    20 component(s) 99.12%   36 component(s) 99.91%
 5 component(s) 88.15%    21 component(s) 99.23%   37 component(s) 99.93%
 6 component(s) 90.9%     22 component(s) 99.33%   38 component(s) 99.94%
 7 component(s) 92.5%     23 component(s) 99.42%   39 component(s) 99.95%
 8 component(s) 93.94%    24 component(s) 99.5%    40 component(s) 99.96%
 9 component(s) 95.18%    25 component(s) 99.56%   41 component(s) 99.97%
10 component(s) 96.22%    26 component(s) 99.62%   42 component(s) 99.97%
11 component(s) 96.79%    27 component(s) 99.67%   43 component(s) 99.98%
12 component(s) 97.2%     28 component(s) 99.71%   44 component(s) 99.99%
13 component(s) 97.58%    29 component(s) 99.75%   45 component(s) 99.99%
14 component(s) 97.92%    30 component(s) 99.79%   46 component(s) 100.0%
15 component(s) 98.18%    31 component(s) 99.82%   47 component(s) 100.0%
16 component(s) 98.41%    32 component(s) 99.84%   48 component(s) 100.0%
```

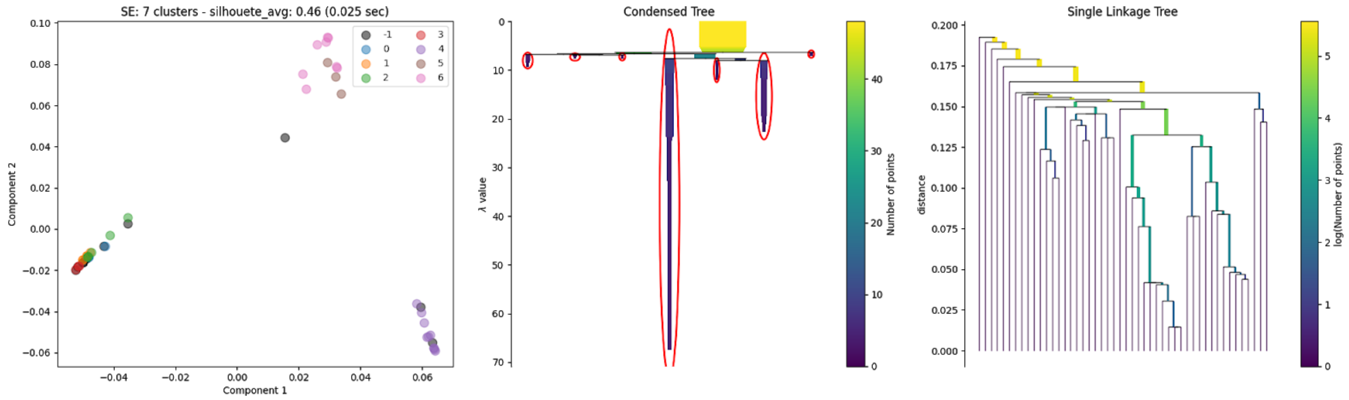
Results

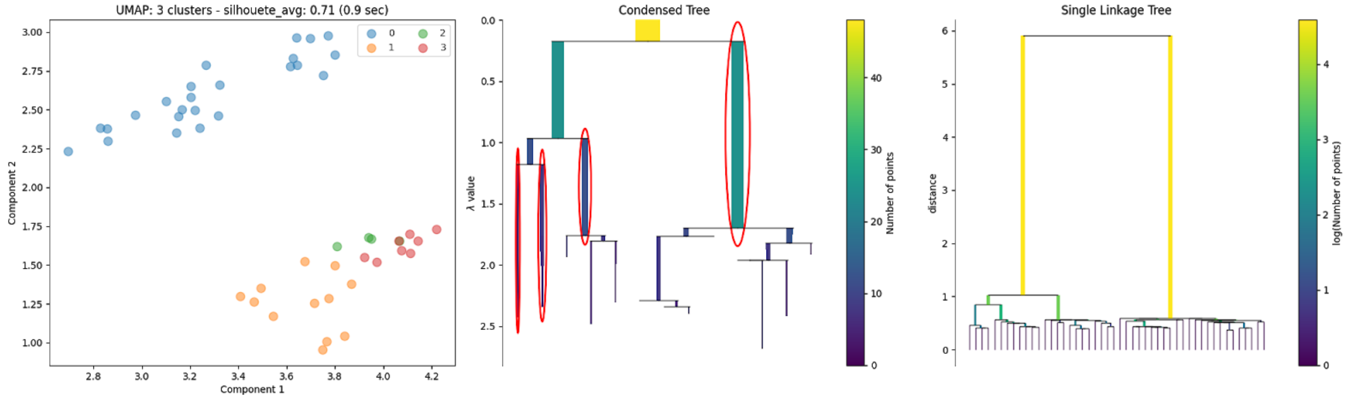
After the first stage, the HDBScan algorithm was applied to all embeddings with the following settings:

model = HDBSCAN(min_cluster_size=2, min_samples=2)

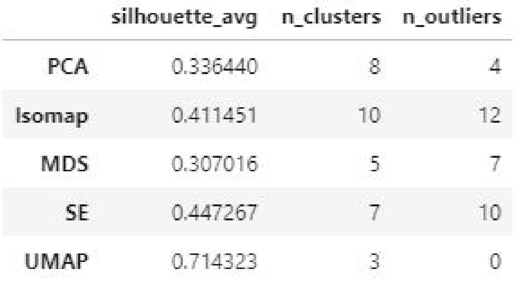
All the results are compiled in the following graphs: (1) shows the cluster numbers (from 0 to n) and the samples that were not clustered are shown in black with the number -1; (2) - The condensed three shows what the cluster hierarchy looks like – which clusters are near each other, or could perhaps be merged, and which are far apart; (3) The single linkage tree also offers a view of how the clusters were formed.





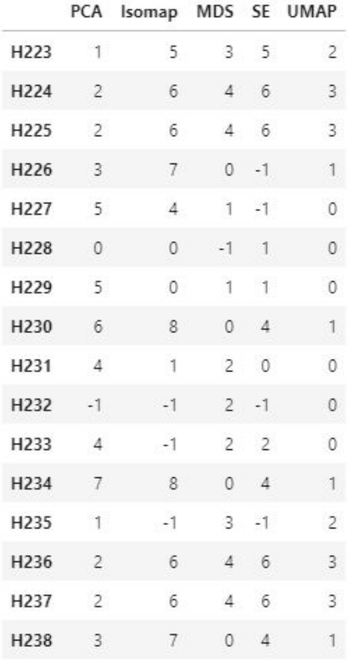


After all the embedding methods were applied and its embedded features were used to cluster the samples, we found the following metrics for each technique:

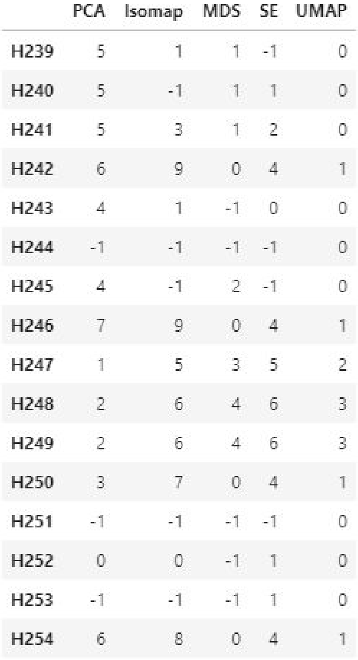


	silhouette_avg	n_clusters	n_outliers
PCA	0.336440	8	4
Isomap	0.411451	10	12
MDS	0.307016	5	7
SE	0.447267	7	10
UMAP	0.714323	3	0

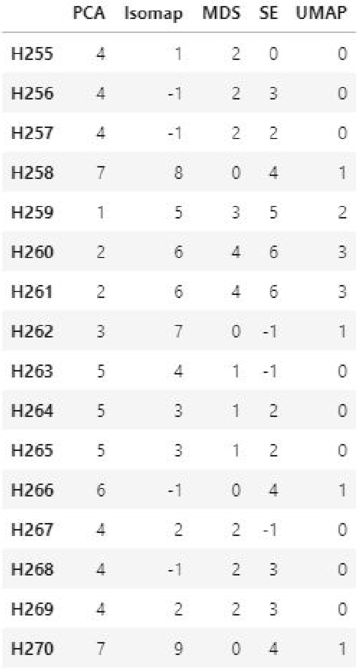
The cluster number assigned to each of the samples, by the HDBScan applied to each of the dimensionality reduction techniques is shown below:



	PCA	Isomap	MDS	SE	UMAP
H223	1	5	3	5	2
H224	2	6	4	6	3
H225	2	6	4	6	3
H226	3	7	0	-1	1
H227	5	4	1	-1	0
H228	0	0	-1	1	0
H229	5	0	1	1	0
H230	6	8	0	4	1
H231	4	1	2	0	0
H232	-1	-1	2	-1	0
H233	4	-1	2	2	0
H234	7	8	0	4	1
H235	1	-1	3	-1	2
H236	2	6	4	6	3
H237	2	6	4	6	3
H238	3	7	0	4	1

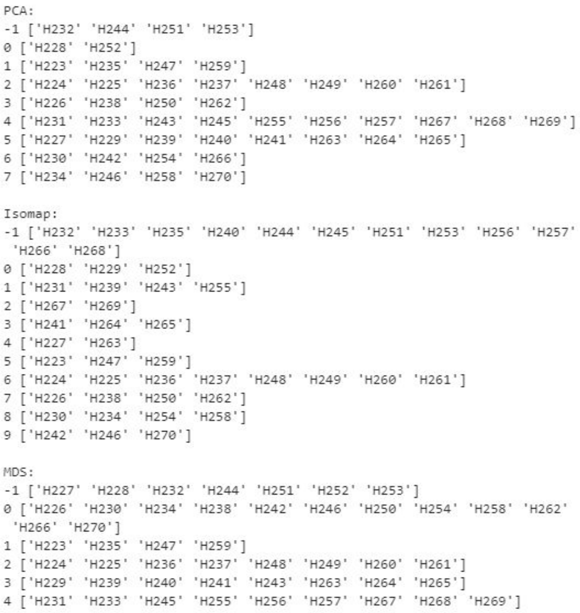


	PCA	Isomap	MDS	SE	UMAP
H239	5	1	1	-1	0
H240	5	-1	1	1	0
H241	5	3	1	2	0
H242	6	9	0	4	1
H243	4	1	-1	0	0
H244	-1	-1	-1	-1	0
H245	4	-1	2	-1	0
H246	7	9	0	4	1
H247	1	5	3	5	2
H248	2	6	4	6	3
H249	2	6	4	6	3
H250	3	7	0	4	1
H251	-1	-1	-1	-1	0
H252	0	0	-1	1	0
H253	-1	-1	-1	1	0
H254	6	8	0	4	1

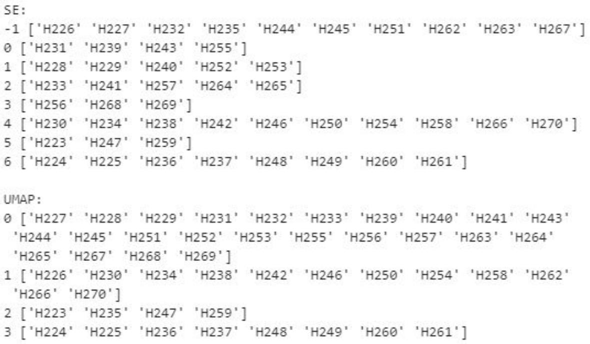


	PCA	Isomap	MDS	SE	UMAP
H255	4	1	2	0	0
H256	4	-1	2	3	0
H257	4	-1	2	2	0
H258	7	8	0	4	1
H259	1	5	3	5	2
H260	2	6	4	6	3
H261	2	6	4	6	3
H262	3	7	0	-1	1
H263	5	4	1	-1	0
H264	5	3	1	2	0
H265	5	3	1	2	0
H266	6	-1	0	4	1
H267	4	2	2	-1	0
H268	4	-1	2	3	0
H269	4	2	2	3	0
H270	7	9	0	4	1

Compiling the results we achieve all the elements in each cluster for all techniques:

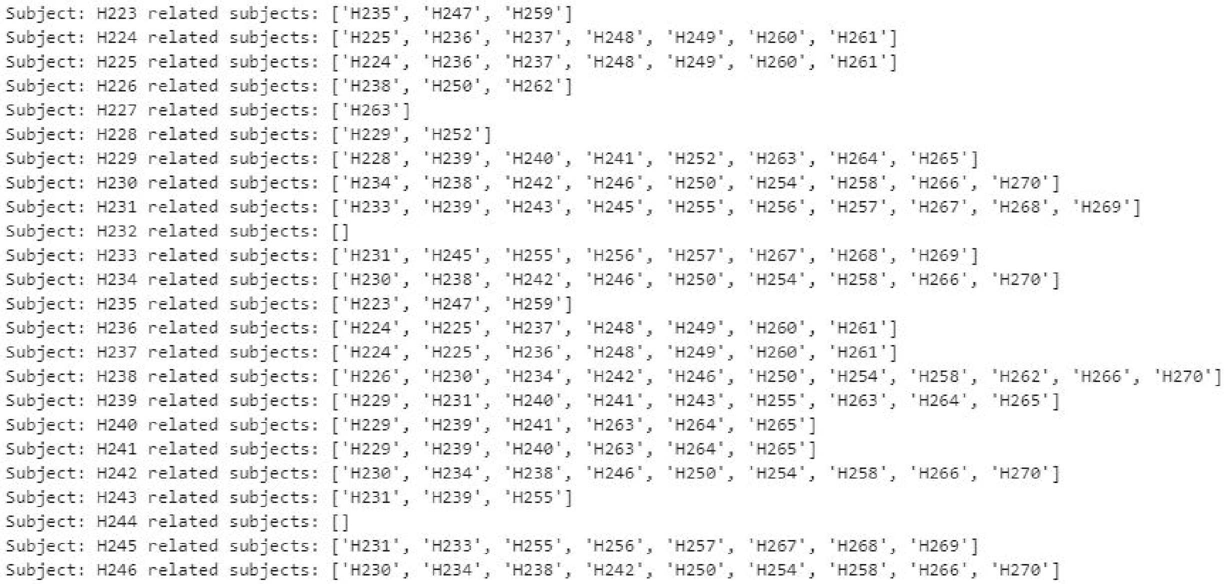


```
PCA:
-1 ['H232' 'H244' 'H251' 'H253']
0 ['H228' 'H252']
1 ['H223' 'H235' 'H247' 'H259']
2 ['H224' 'H225' 'H236' 'H237' 'H248' 'H249' 'H260' 'H261']
3 ['H226' 'H238' 'H250' 'H262']
4 ['H231' 'H233' 'H243' 'H245' 'H255' 'H256' 'H257' 'H267' 'H268' 'H269']
5 ['H227' 'H229' 'H239' 'H240' 'H241' 'H263' 'H264' 'H265']
6 ['H230' 'H242' 'H254' 'H266']
7 ['H234' 'H246' 'H258' 'H270']

Isomap:
-1 ['H232' 'H233' 'H235' 'H240' 'H244' 'H245' 'H251' 'H253' 'H256' 'H257'
 'H266' 'H268']
0 ['H228' 'H229' 'H252']
1 ['H231' 'H239' 'H243' 'H255']
2 ['H267' 'H269']
3 ['H241' 'H264' 'H265']
4 ['H227' 'H263']
5 ['H223' 'H247' 'H259']
6 ['H224' 'H225' 'H236' 'H237' 'H248' 'H249' 'H260' 'H261']
7 ['H226' 'H238' 'H250' 'H262']
8 ['H230' 'H234' 'H254' 'H258']
9 ['H242' 'H246' 'H270']

MDS:
-1 ['H227' 'H228' 'H232' 'H244' 'H251' 'H252' 'H253']
0 ['H226' 'H230' 'H234' 'H238' 'H242' 'H246' 'H250' 'H254' 'H258' 'H262'
 'H266' 'H270']
1 ['H223' 'H235' 'H247' 'H259']
2 ['H224' 'H225' 'H236' 'H237' 'H248' 'H249' 'H260' 'H261']
3 ['H229' 'H239' 'H240' 'H241' 'H243' 'H263' 'H264' 'H265']
4 ['H231' 'H233' 'H245' 'H255' 'H256' 'H257' 'H267' 'H268' 'H269']
```

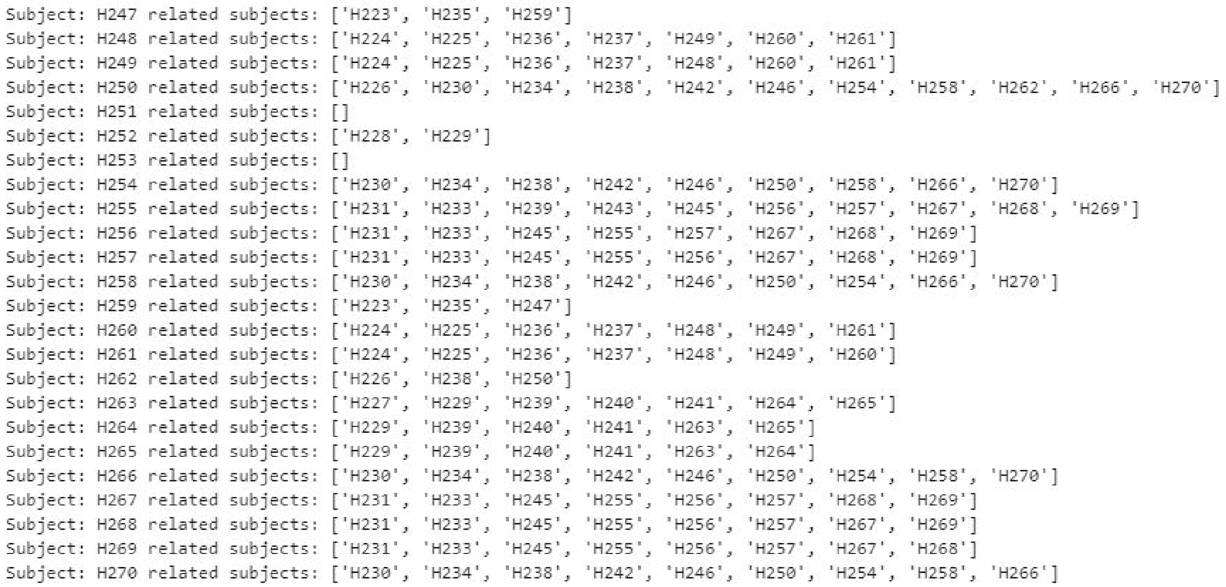


```
SE:
-1 ['H226' 'H227' 'H232' 'H235' 'H244' 'H245' 'H251' 'H262' 'H263' 'H267']
0 ['H231' 'H239' 'H243' 'H255']
1 ['H228' 'H229' 'H240' 'H252' 'H253']
2 ['H233' 'H241' 'H257' 'H264' 'H265']
3 ['H256' 'H268' 'H269']
4 ['H230' 'H234' 'H238' 'H242' 'H246' 'H250' 'H254' 'H258' 'H266' 'H270']
5 ['H223' 'H247' 'H259']
6 ['H224' 'H225' 'H236' 'H237' 'H248' 'H249' 'H260' 'H261']

UMAP:
0 ['H227' 'H228' 'H229' 'H231' 'H232' 'H233' 'H239' 'H240' 'H241' 'H243'
 'H244' 'H245' 'H251' 'H252' 'H253' 'H255' 'H256' 'H257' 'H263' 'H264'
 'H265' 'H267' 'H268' 'H269']
1 ['H226' 'H230' 'H234' 'H238' 'H242' 'H246' 'H250' 'H254' 'H258' 'H262'
 'H266' 'H270']
2 ['H223' 'H235' 'H247' 'H259']
3 ['H224' 'H225' 'H236' 'H237' 'H248' 'H249' 'H260' 'H261']
```

From the previous results we can find, from each subject's point of view, which other subjects are part of the same cluster most of the time (at least in 3 out of 5). Also, the ones that are classified as “-1” more than 3 times are considered to not be related to any other subject in the database.

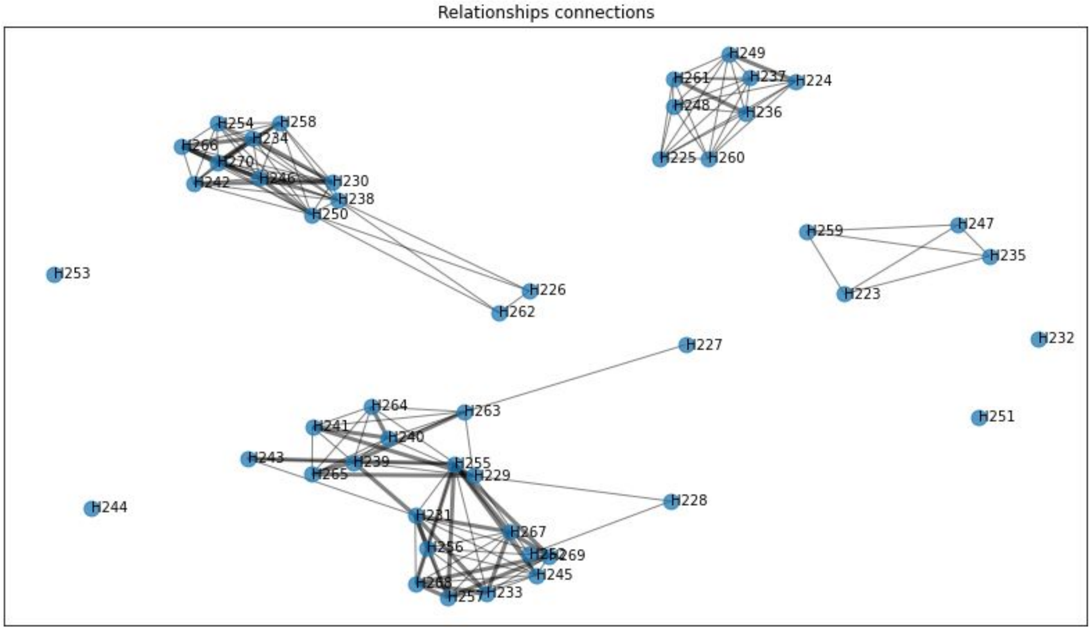


```
Subject: H223 related subjects: ['H235', 'H247', 'H259']
Subject: H224 related subjects: ['H225', 'H236', 'H237', 'H248', 'H249', 'H260', 'H261']
Subject: H225 related subjects: ['H224', 'H236', 'H237', 'H248', 'H249', 'H260', 'H261']
Subject: H226 related subjects: ['H238', 'H250', 'H262']
Subject: H227 related subjects: ['H263']
Subject: H228 related subjects: ['H229', 'H252']
Subject: H229 related subjects: ['H228', 'H239', 'H240', 'H241', 'H252', 'H263', 'H264', 'H265']
Subject: H230 related subjects: ['H234', 'H238', 'H242', 'H246', 'H250', 'H254', 'H258', 'H266', 'H270']
Subject: H231 related subjects: ['H233', 'H239', 'H243', 'H245', 'H255', 'H256', 'H257', 'H267', 'H268', 'H269']
Subject: H232 related subjects: []
Subject: H233 related subjects: ['H231', 'H245', 'H255', 'H256', 'H257', 'H267', 'H268', 'H269']
Subject: H234 related subjects: ['H230', 'H238', 'H242', 'H246', 'H250', 'H254', 'H258', 'H266', 'H270']
Subject: H235 related subjects: ['H223', 'H247', 'H259']
Subject: H236 related subjects: ['H224', 'H225', 'H237', 'H248', 'H249', 'H260', 'H261']
Subject: H237 related subjects: ['H224', 'H225', 'H236', 'H248', 'H249', 'H260', 'H261']
Subject: H238 related subjects: ['H226', 'H230', 'H234', 'H242', 'H246', 'H250', 'H254', 'H258', 'H262', 'H266', 'H270']
Subject: H239 related subjects: ['H229', 'H231', 'H240', 'H241', 'H243', 'H255', 'H263', 'H264', 'H265']
Subject: H240 related subjects: ['H229', 'H239', 'H241', 'H263', 'H264', 'H265']
Subject: H241 related subjects: ['H229', 'H239', 'H240', 'H263', 'H264', 'H265']
Subject: H242 related subjects: ['H230', 'H234', 'H238', 'H246', 'H250', 'H254', 'H258', 'H266', 'H270']
Subject: H243 related subjects: ['H231', 'H239', 'H255']
Subject: H244 related subjects: []
Subject: H245 related subjects: ['H231', 'H233', 'H255', 'H256', 'H257', 'H267', 'H268', 'H269']
Subject: H246 related subjects: ['H230', 'H234', 'H238', 'H242', 'H250', 'H254', 'H258', 'H266', 'H270']
```



```
Subject: H247 related subjects: ['H223', 'H235', 'H259']
Subject: H248 related subjects: ['H224', 'H225', 'H236', 'H237', 'H249', 'H260', 'H261']
Subject: H249 related subjects: ['H224', 'H225', 'H236', 'H237', 'H248', 'H260', 'H261']
Subject: H250 related subjects: ['H226', 'H230', 'H234', 'H238', 'H242', 'H246', 'H254', 'H258', 'H262', 'H266', 'H270']
Subject: H251 related subjects: []
Subject: H252 related subjects: ['H228', 'H229']
Subject: H253 related subjects: []
Subject: H254 related subjects: ['H230', 'H234', 'H238', 'H242', 'H246', 'H250', 'H258', 'H266', 'H270']
Subject: H255 related subjects: ['H231', 'H233', 'H239', 'H243', 'H245', 'H256', 'H257', 'H267', 'H268', 'H269']
Subject: H256 related subjects: ['H231', 'H233', 'H245', 'H255', 'H257', 'H267', 'H268', 'H269']
Subject: H257 related subjects: ['H231', 'H233', 'H245', 'H255', 'H256', 'H267', 'H268', 'H269']
Subject: H258 related subjects: ['H230', 'H234', 'H238', 'H242', 'H246', 'H250', 'H254', 'H266', 'H270']
Subject: H259 related subjects: ['H223', 'H235', 'H247']
Subject: H260 related subjects: ['H224', 'H225', 'H236', 'H237', 'H248', 'H249', 'H261']
Subject: H261 related subjects: ['H224', 'H225', 'H236', 'H237', 'H248', 'H249', 'H260']
Subject: H262 related subjects: ['H226', 'H238', 'H250']
Subject: H263 related subjects: ['H227', 'H229', 'H239', 'H240', 'H241', 'H264', 'H265']
Subject: H264 related subjects: ['H229', 'H239', 'H240', 'H241', 'H263', 'H265']
Subject: H265 related subjects: ['H229', 'H239', 'H240', 'H241', 'H263', 'H264']
Subject: H266 related subjects: ['H230', 'H234', 'H238', 'H242', 'H246', 'H250', 'H254', 'H258', 'H270']
Subject: H267 related subjects: ['H231', 'H233', 'H245', 'H255', 'H256', 'H257', 'H268', 'H269']
Subject: H268 related subjects: ['H231', 'H233', 'H245', 'H255', 'H256', 'H257', 'H267', 'H269']
Subject: H269 related subjects: ['H231', 'H233', 'H245', 'H255', 'H256', 'H257', 'H267', 'H268']
Subject: H270 related subjects: ['H230', 'H234', 'H238', 'H242', 'H246', 'H250', 'H254', 'H258', 'H266']
```

From these results, we see that in most cases, if A is related to B, also, B is related to A. From our results, this is not always true: In some cases, we would say that A is related to B, but we’re not sure that B is related to A. The following graph shows the relationships between all subjects. The ones that have a simple one-sided connection are shown with narrower edges and the ones with two-sided connections are shown with thicker edges.



Conclusion

The ensemble cluster is a relatively easy-to-implement technique, and the results seem to be reasonable at least. The main difficulties found in this work were due to the lack of information from the database and the lack of knowledge about genomic data. The choice for the HDBScan algorithm seems to be reasonable also since it can find the more stable clusters and decide the number of clusters without the need for the user to do so.