Test results

Milica Simić

November 2023

Content

Data

- Data used in this task: Mouse embryo brain
- Data contains 59704 cells and 4 columns with values:
  - cell name (`cell_id`)
  - spatial coordinates (`x, y`)
  - cell type (`sim anno`)

	cell_ID	x	y	sim anno
0	CELL.61	-5591.0	-1156.0	Ery
1	CELL.84	-5538.0	-1185.0	Ery
2	CELL.105	-5496.0	-1206.0	Ery
3	CELL.131	-5519.0	-1235.0	Ery
4	CELL.132	-5546.0	-1235.0	Ery
...	...	...	...	...
59699	CELL.70553	-1463.0	-8763.0	DorsHb RGC
59700	CELL.70555	-1175.0	-8770.0	Mixed GlioB
59701	CELL.70556	-1365.0	-8764.0	Mixed GlioB
59702	CELL.70557	-1425.0	-8776.0	DorsHb RGC
59703	CELL.70558	-1396.0	-8774.0	DorsHb RGC

59704 rows × 4 columns

Data

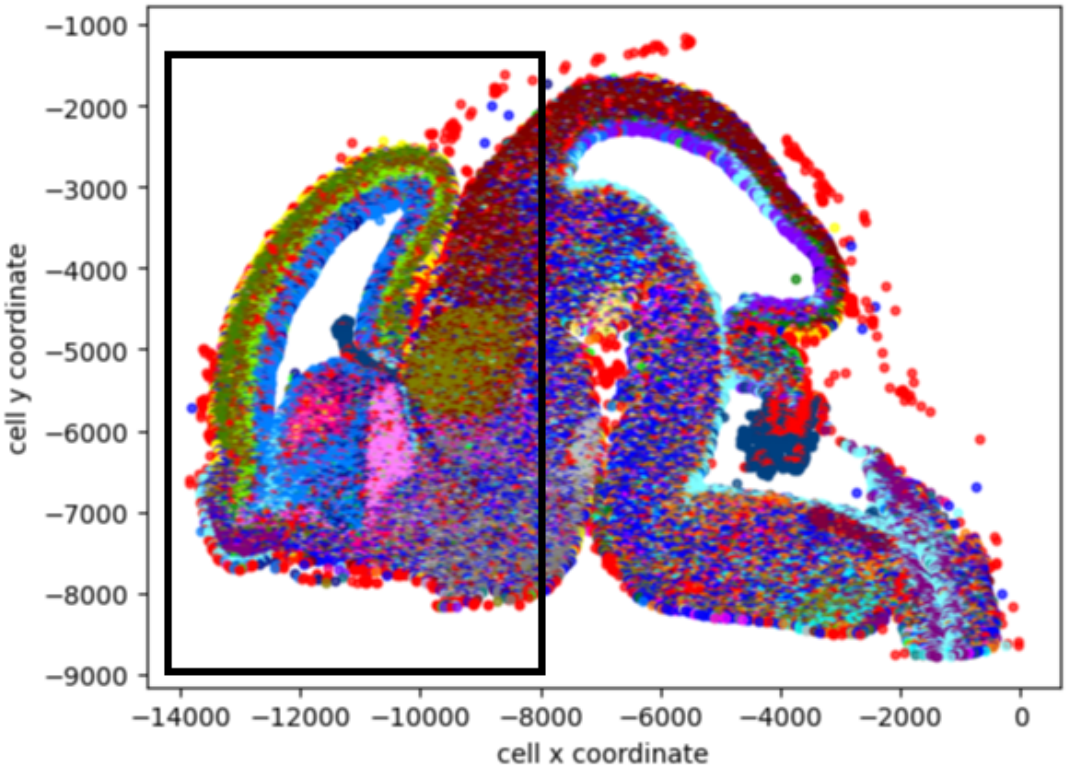
- Reduction of data was necessary due to memory and time constraints
- The data was filtered to remove all cells with values for x greater than -8000
- Reduced data contains 31112 cells

	cell_ID	x	y	sim anno
0	CELL.523	-8598.0	-1609.0	Ery
1	CELL.662	-8135.0	-1686.0	Ery
2	CELL.918	-8628.0	-1775.0	Ery
3	CELL.933	-8745.0	-1779.0	Ery
4	CELL.942	-8786.0	-1779.0	Ery
...	...	...	...	...
31107	CELL.68191	-9542.0	-8167.0	Ery
31108	CELL.68234	-9631.0	-8172.0	Hb Glu NeuB
31109	CELL.68264	-9610.0	-8173.0	Mb RGC
31110	CELL.68268	-9493.0	-8171.0	Ery
31111	CELL.68334	-9395.0	-8176.0	Mb RGC

31112 rows × 4 columns

Data

- Reduced data is contained within the black rectangle

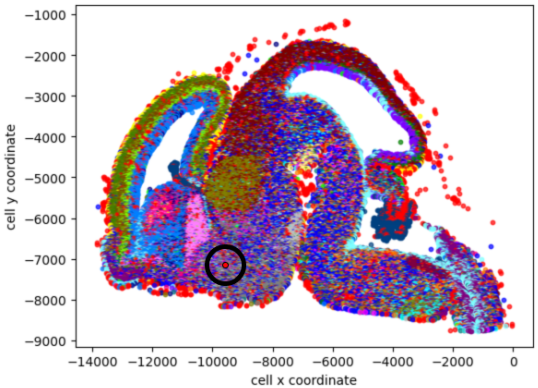


Task

- In the context of tissue analysis, it is important to recognize that every cell within a tissue interacts with its neighboring cells, typically performing specific functions.
- Consequently, cells sharing similar percentage of cell types in their surroundings can be identified as spatial community
- The objectives of the task:
  1. Categorize cells into spatial communities by evaluating the percentage of cell types found in their immediate surroundings
  2. Calculate for each community the mixture (count and percentage) of cell types that are present in it
  3. Calculate the homogeneity score for each community and decide whether the community is homogeneous (community with very similar percentages of cell types in all of its parts) or heterogeneous (community in which percentages of cell types vary significantly across different parts of it)

Calculation of the percentage of cell types in cell's neighborhood

- Input: `Cell cell` (red dot on the image below) and `integer neighborhood_radius` (radius of black circle on the image below)
- Output: `list types_percentage`, where `types_percentage[k]` represents the percentage of cells in the neighborhood of the cell (indicated by the black circle in the image below) that belong to the cell type with number `k`. All cell types in the tissue are mapped to integers from 0 to `num_of_types - 1`



- After this step, we have a list of the percentage of cell types in each cell's neighborhood, allowing us to compare them based on this information

Comparing cells based on the percentage of cell types in their neighborhood

- Distance functions used: Manhattan, Euclidean and Hamming
- For the Hamming distance, there is a parameter called `hamming_param`. For example, if `types_percentage = [0.1, 0.5, 0.4]` and `hamming_param = 0.3`, then the modified percentages will be `[0, 1, 1]`. In this case, only the first element (0.1) is less than 0.3, so it will be set to 0, and the rest will be set to 1. And these modified percentages of cells will be compared when comparing two cells using the Hamming distance.

Clustering cells based on the similarity of cell types present in their neighborhood.

- Clustering algorithm: Hierarchical clustering (Agglomerative) and Leiden clustering
- Only the results for the Agglomerative clustering will be presented, as it is better suited for the current problem
- For Agglomerative clustering there are two optional parameters:
  - `linkage_method` (single, average, ward, centroid) where default value is single
  - `threshold` where default value is None, and in such cases, the optimal threshold will be calculated using the Silhouette method
- Agglomerative clustering will be performed on the distance matrix `percentage_distance_matrix`, where `percentage_distance_matrix[cell_i.id_num][cell_j.id_num] = dist(cell_i.types_percentage, cell_i.types_percentage)`, and the distance function `dist` can be one of Manhattan, Euclidean, or Hamming distances

Adjustable parameters

- In summary, the adjustable parameters are:
  - `neighborhood_radius` (mandatory parameter)
  - `distance_function` (Manhattan which is default, Euclidean and Hamming with parameter)
  - `linkage_method` (single which is default, average, ward and centroid) and `threshold` (default is None) for Agglomerative clustering

Result 1

- Results where:
  - `neighborhood_radius = 200` (the radius that corresponds to the red circle in the image below)
  - `distance_function = Manhattan`
  - `linkage_method = average`

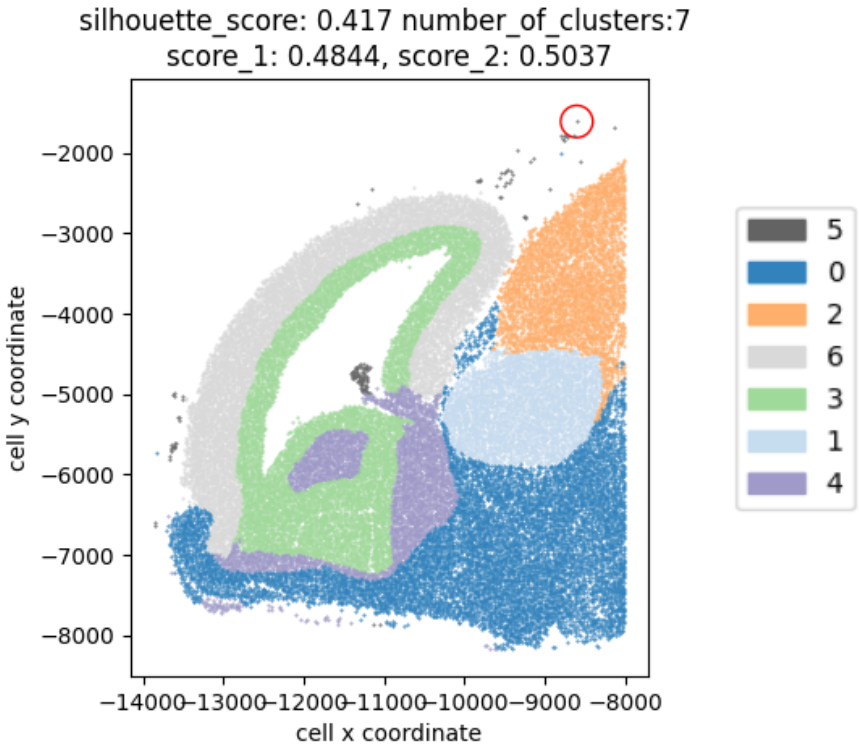


Figure: Clusters

Result 1 - Clusters statistics

- The mean of distribution of percentage distances between all cells is 1.48, as shown in the image below

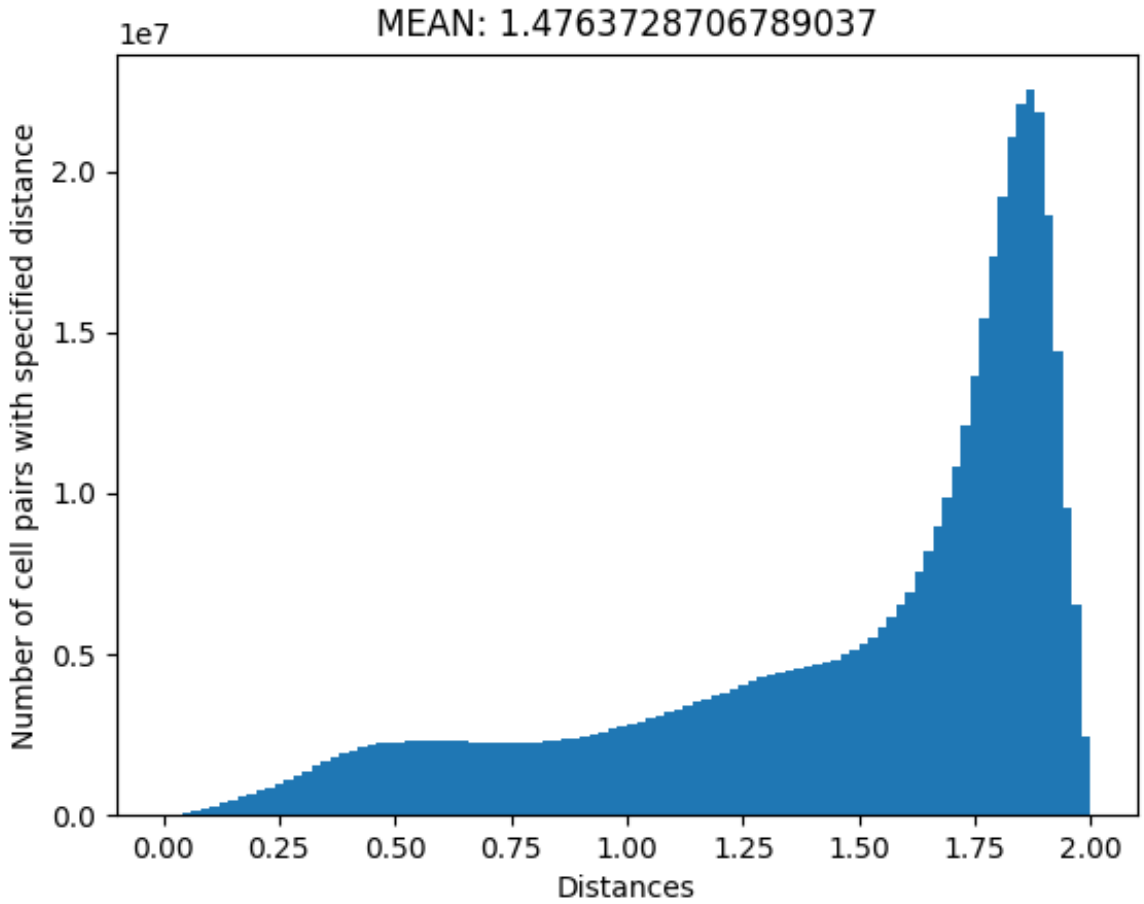


Figure: Distribution of percentage distances between all cells

Result 1 - Homogeneous or Heterogeneous clusters

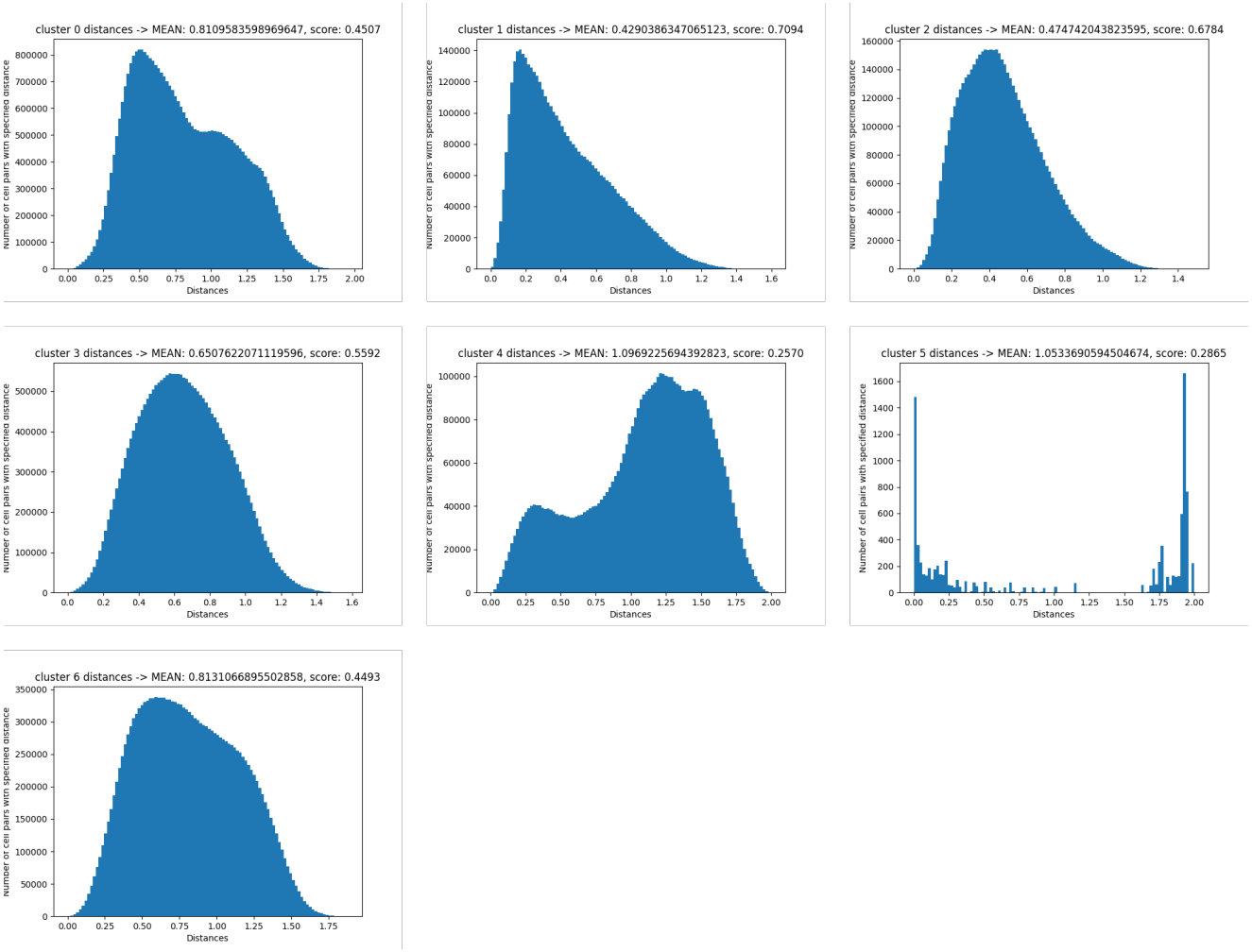


Figure: Distribution of percentage distances between cells for each cluster

Result 1 - Homogeneous or Heterogeneous clusters

- In the image above, distributions of percentage distances between cells for each cluster are displayed
- This was done to determine whether a cluster is homogeneous or heterogeneous. If a cluster is homogeneous, there should be a significant number of percentage distances between cells that are close to zero as this indicates that neighborhood of those cells are similar
- If there is a significant number of homogeneous clusters, it indicates the effectiveness of the clustering, as our goal is to group cells with very similar surroundings.

Cluster homogeneity score

- Cluster c homogeneity score

$$HS(c) = \begin{cases} 1 - m(c)/M & , \text{ if m(c)} < \text{M} \\ 0 & , \text{ otherwise} \end{cases}$$

  $m(c)$ - mean of distribution of percentage distances between cells in the cluster c
  M - mean of distribution of percentage distances between all cells in the tissue

- Cluster homogeneity score takes values from 0 to 1 meaning:
  - if $HS(c) > 0.5$, then cluster c is homogeneous
  - otherwise, cluster c is heterogeneous

Total homogeneity score

- Total homogeneity score of clustering C

$$THS(C) = \frac{\sum_{c \in C} HS(c) \cdot n(c)}{N}$$

$n(c)$ - number of cells in the cluster c
N - number of cells in the tissue

- Total homogeneity score takes values from 0 to 1 meaning: value close to 1 indicates that the clustering is of high quality, while a value close to 0 suggests that the clustering is of poor quality.

Result 1 - Homogeneous or Heterogeneous clusters

- From the image above, we can observe that we have 7 clusters from which 1, 2, and 3 are homogeneous, and cluster 0 and 6 are close to homogeneous
- Furthermore, clusters 4 and 5 have the lowest homogeneity score and are heterogeneous clusters
- Total homogeneity score of this clustering is 0.5 and silhouette score is 0.42

Result 1 (statistics) - Number of different cell types in each cluster

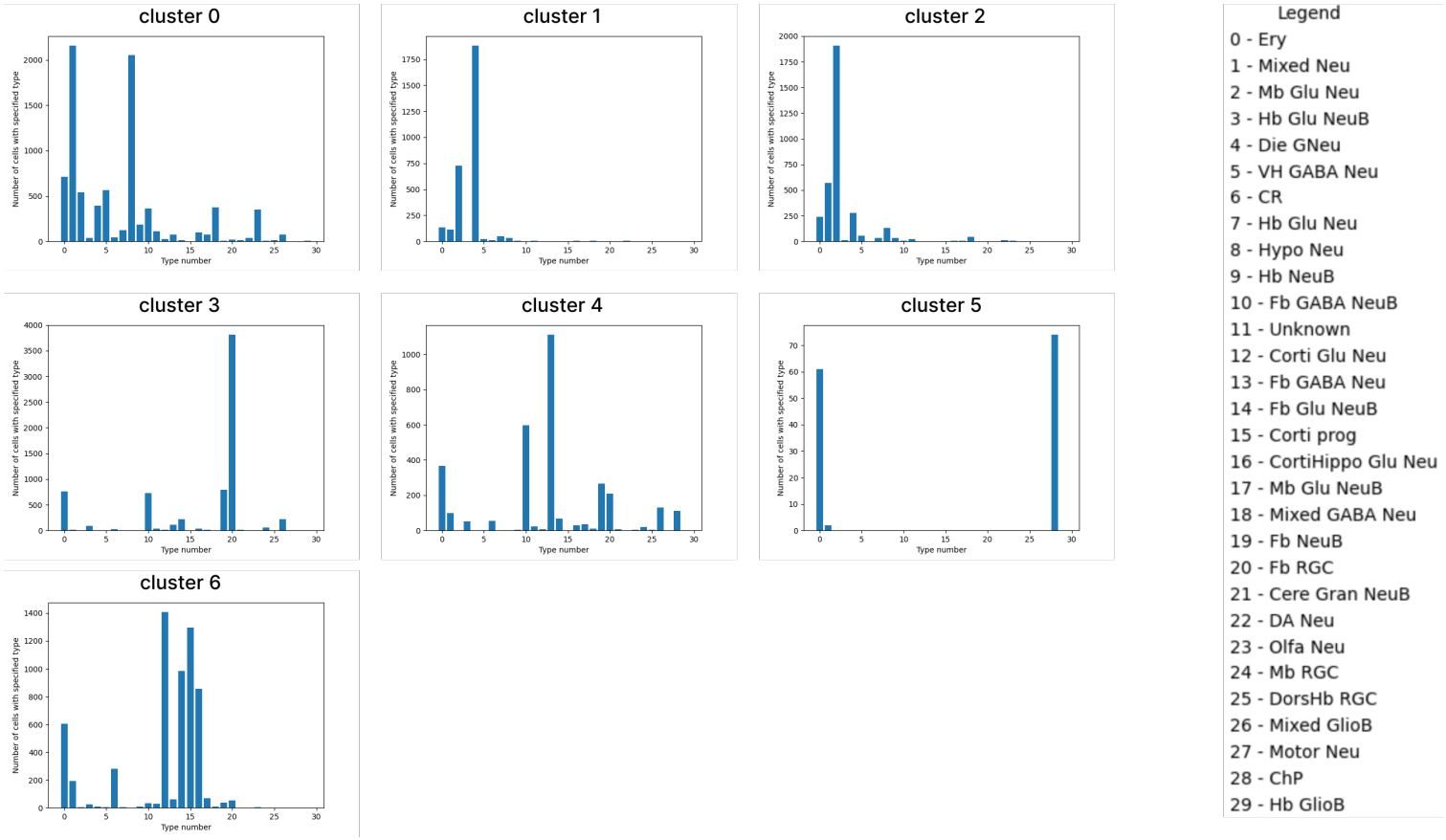


Figure: Number of different cell types in each cluster

Result 1 (statistics) - Percentage of different cell types in each cluster

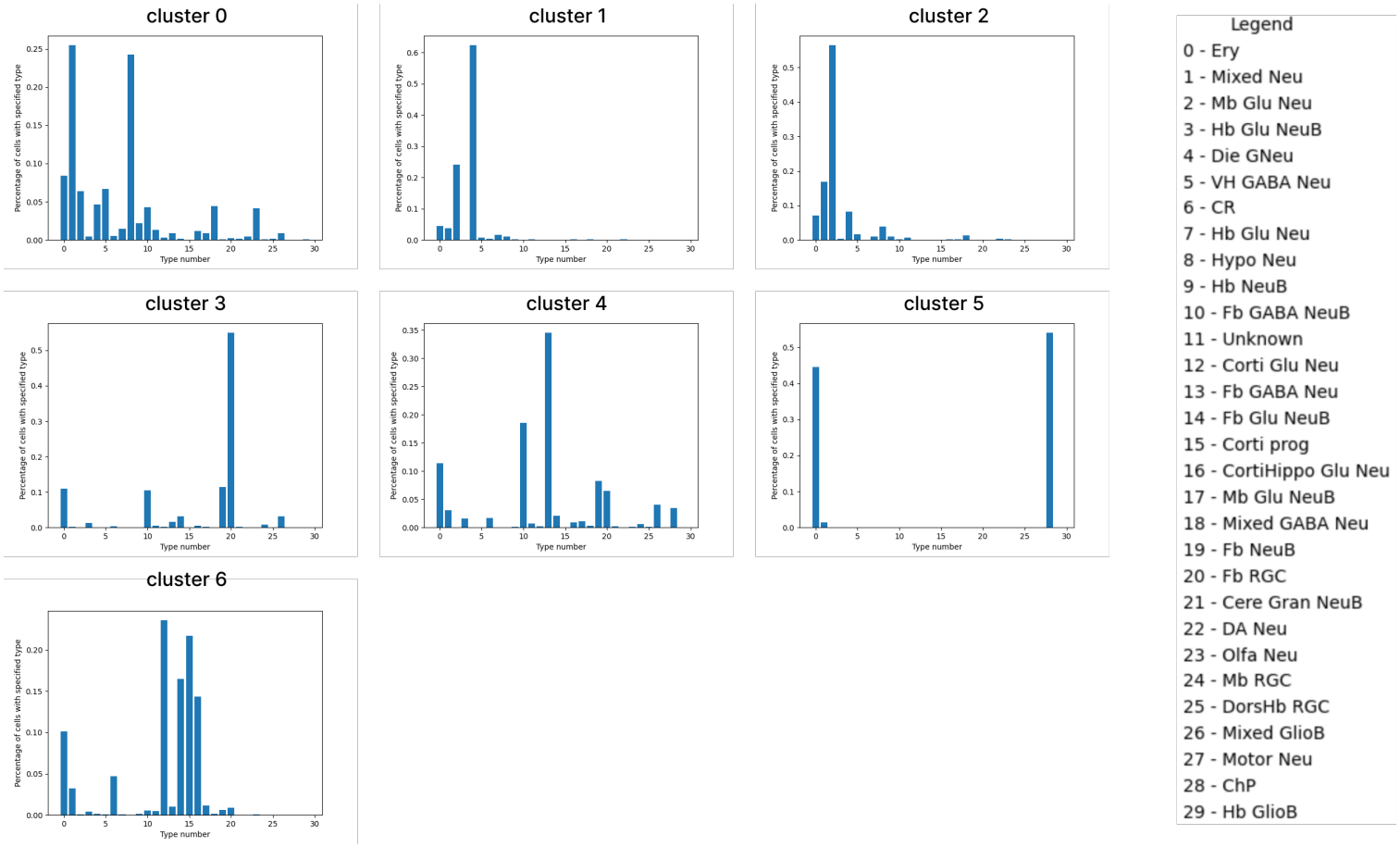


Figure: Percentage of different cell types in each cluster

Result 1 - Decreasing the threshold

- Decreasing the threshold (from 68 to 55) did result in a higher number of clusters with slight improvement in homogeneity score
- The gray cluster (cluster 6) from the previous clustering has now been divided into clusters 13 and 14. This division has resulted in higher homogeneity scores for clusters 13 (0.62) and 14 (0.6) than what was observed for the single cluster 6 (0.45) in the previous clustering.

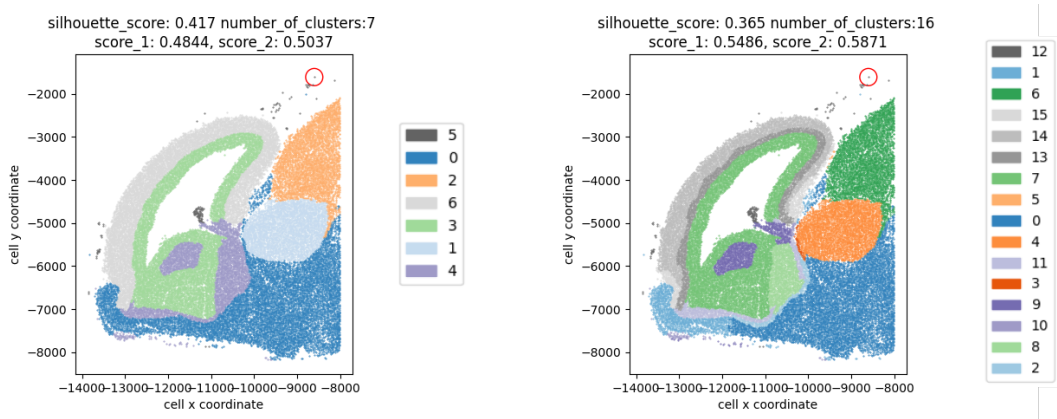


Figure: Clusters with `threshold` 64 and 55, respectively

Result 1 - Decreasing the threshold

- The purple cluster (cluster 4) from the previous clustering has now been divided into clusters 8 , 9 and 11 .This division has resulted in higher homogeneity scores for clusters 8 (0.64), 9 (0.45) and 11 (0.46) than what was observed for the single cluster 4 (0.26) in the previous clustering.
- The blue cluster (cluster 0) from the previous clustering has now been divided into clusters 0, 1 and 2.This division has resulted in higher homogeneity scores for clusters 0 (0.56), and 2 (0.54) and lower for cluster 1 (0.39) than what was observed for the single cluster 0 (0.45) in the previous clustering.

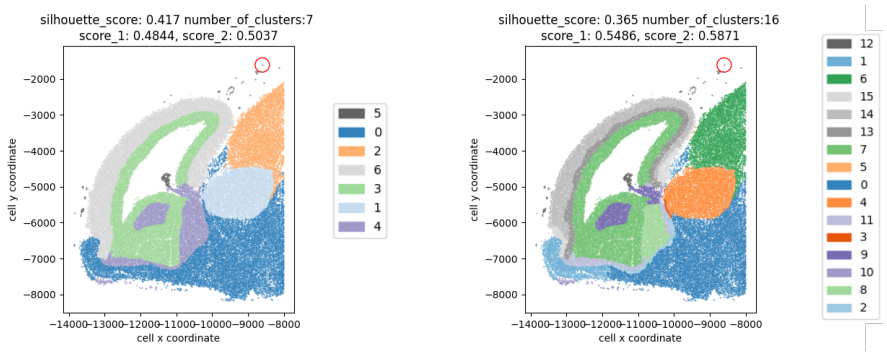


Figure: Clusters with `threshold` 68 and 55, respectively

Result 1 - Decreasing the threshold

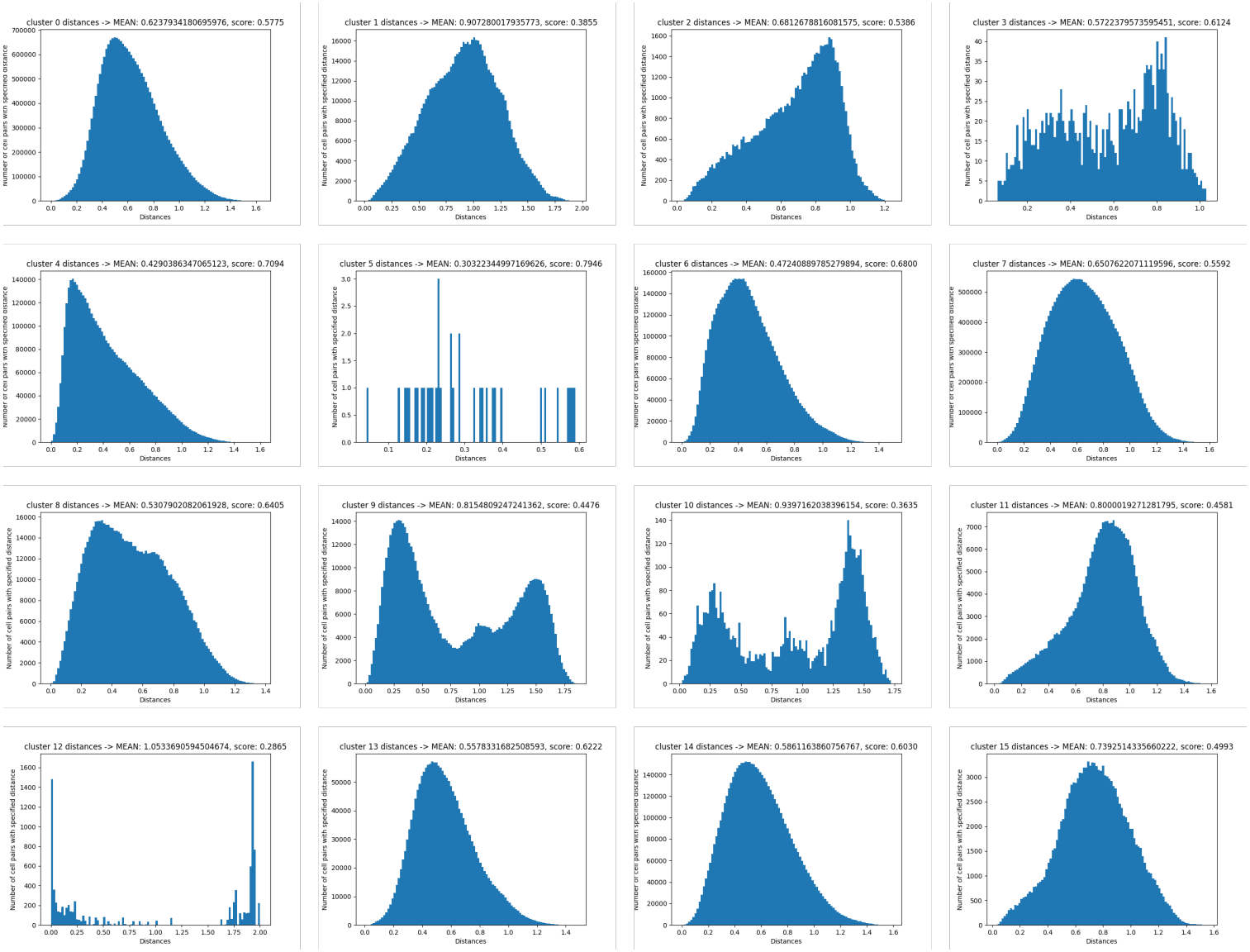


Figure: Distribution of distances between cells for each cluster

Result 1 - Decreasing the threshold

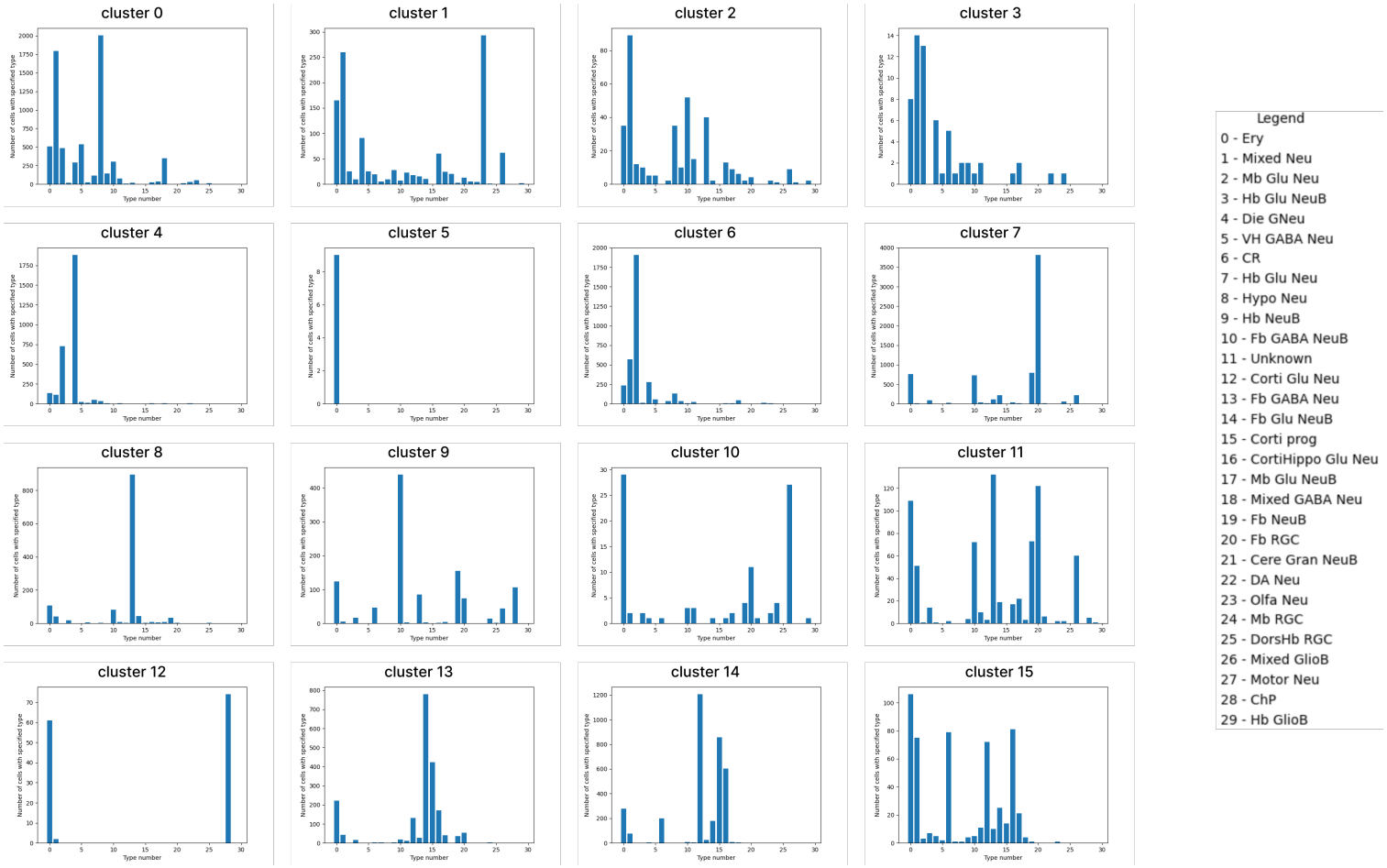


Figure: Number of different cell types in each cluster

Result 1 - Decreasing the threshold

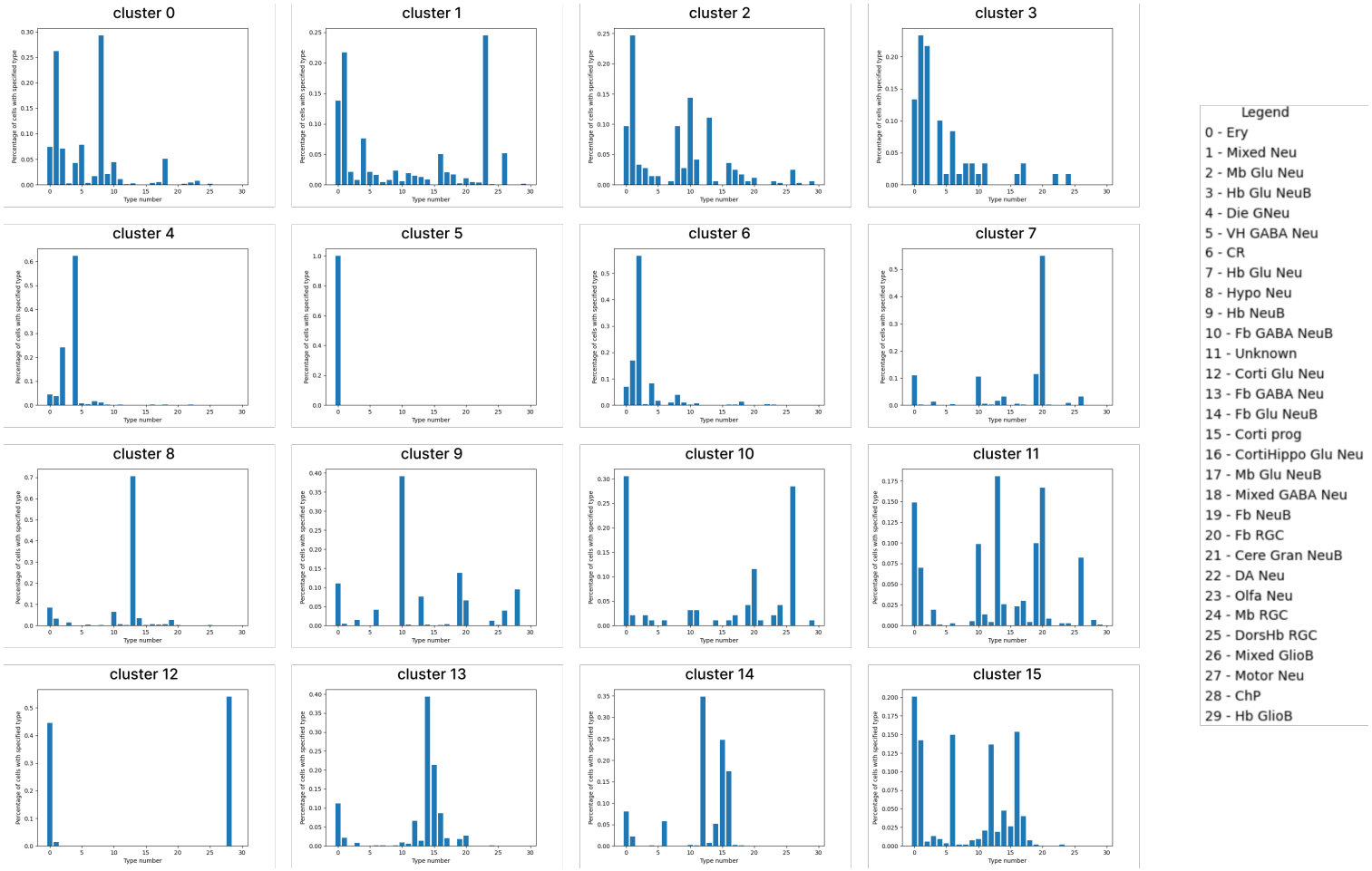


Figure: Percentage of different cell types in each cluster

Result 1 - Decreasing the neighborhood radius

- Increasing the neighborhood (from 200 to 100) radius did result in higher number of clusters, with a slightly lower silhouette and total homogeneity score.
- The purple cluster (cluster 4) from the previous clustering has now been divided into clusters 5 and 6 .This division has resulted in higher homogeneity scores for clusters 5 (0.59) and 6 (0.36) than what was observed for the single cluster 4 (0.26) in the previous clustering.

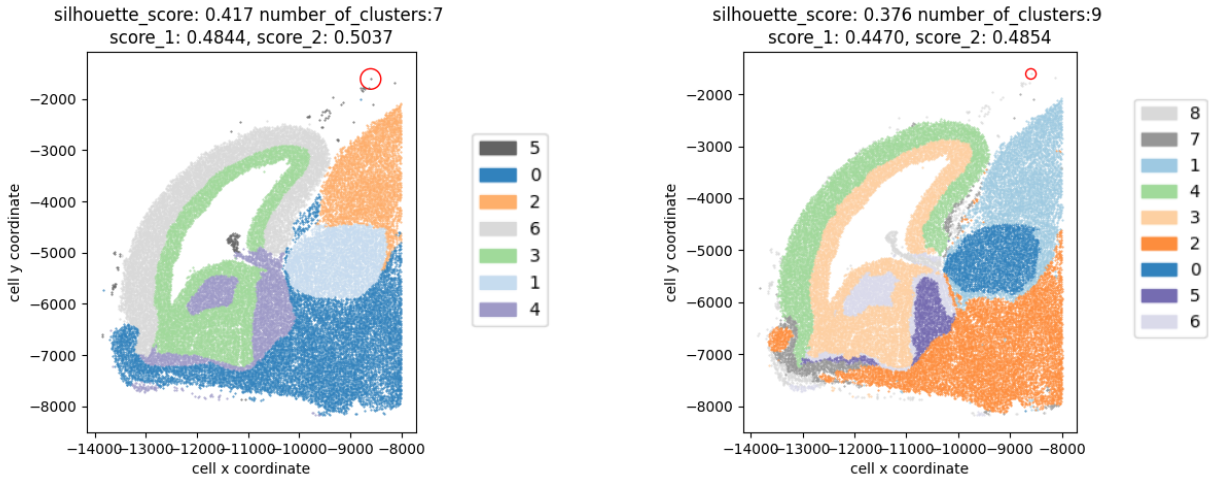


Figure: Clusters with `neighborhood_radius` 200 and 100, respectively

Result 1 - Decreasing the neighborhood radius

- The blue cluster (cluster 0) from the previous clustering has now been divided into clusters 2 and 7.This division has resulted in higher homogeneity scores for clusters 2 (0.47) and lower for cluster 7 (0.3) than what was observed for the single cluster 0 (0.45) in the previous clustering.
- The green cluster (cluster 3) from the previous clustering (0.56) has slightly higher homogeneity scores than cluster 3 in new clustering (0.51), same as clusters 6 from the previous clustering (0.45) and cluster 4 from the new clustering (0.37).

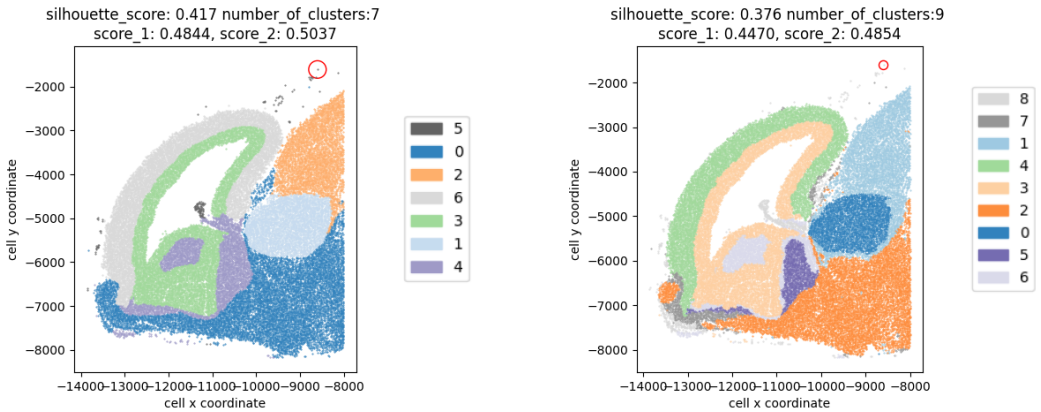


Figure: Clusters with `neighborhood_radius` 200 and 100, respectively

Result 1 - Decreasing the neighborhood radius

- The mean of distribution of percentage distances between all cells is 1.88, as shown in the image below

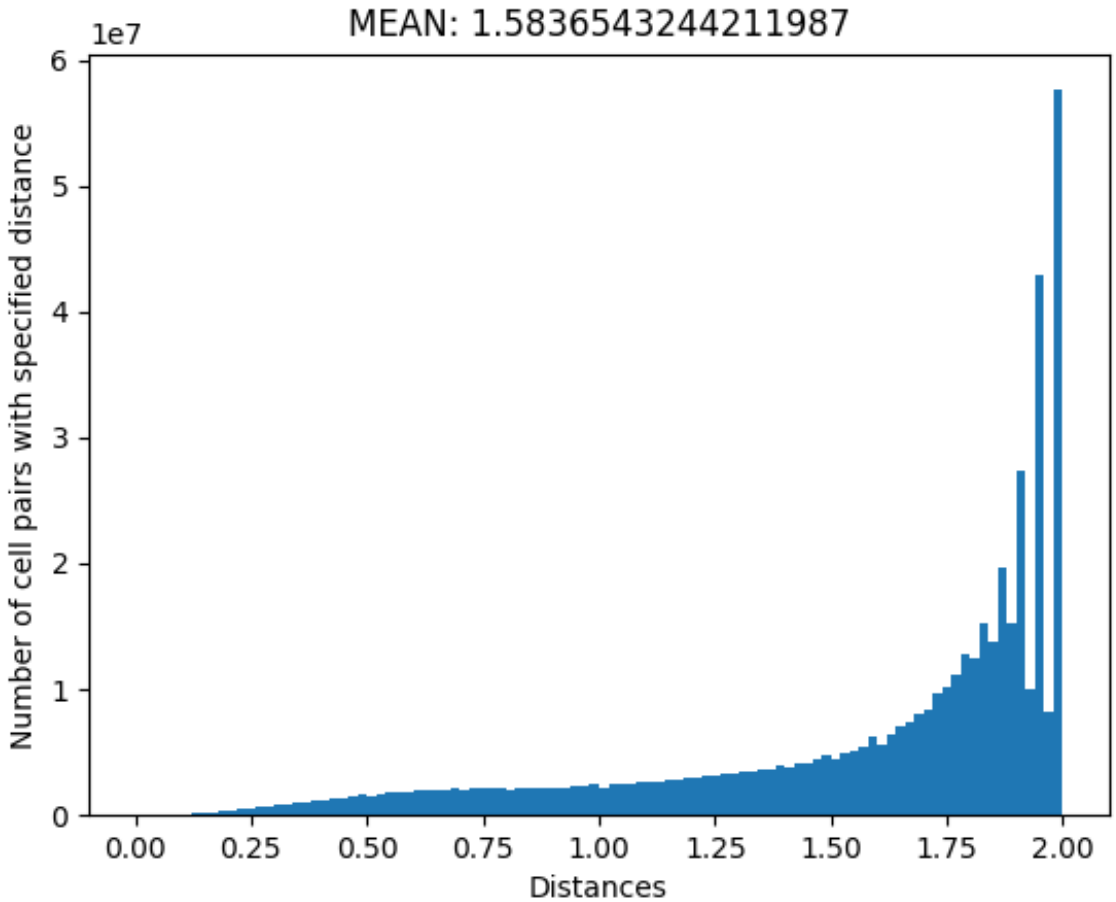


Figure: Distribution of percentage distances between all cells

Result 1 - Decreasing the neighborhood radius

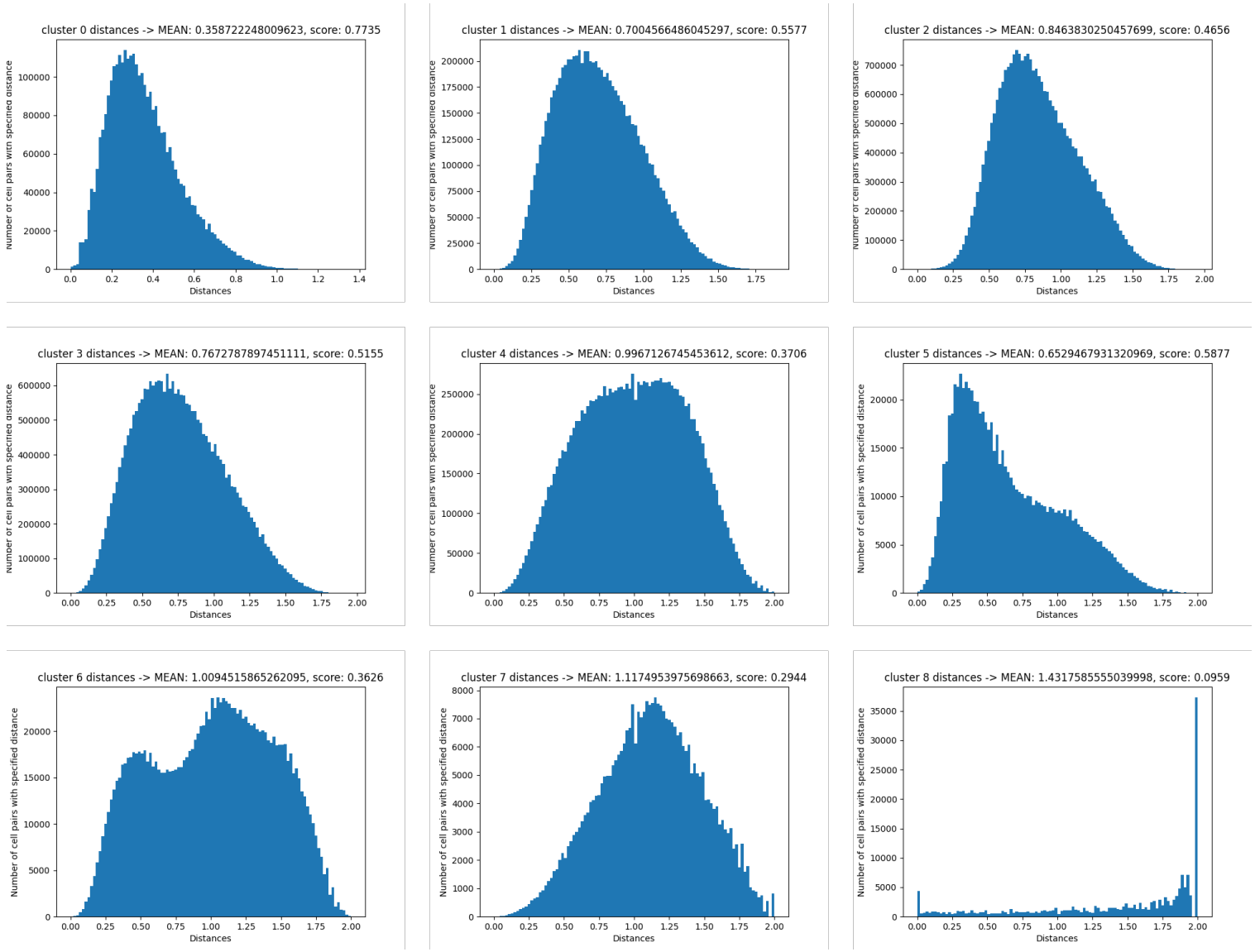


Figure: Distribution of distances between cells for each cluster

Result 2

- Results where:
  - `neighborhood_radius = 200` (the radius that corresponds to the red circle in the image below)
  - `distance_function = Euclidean`
  - `linkage_method = centroid`

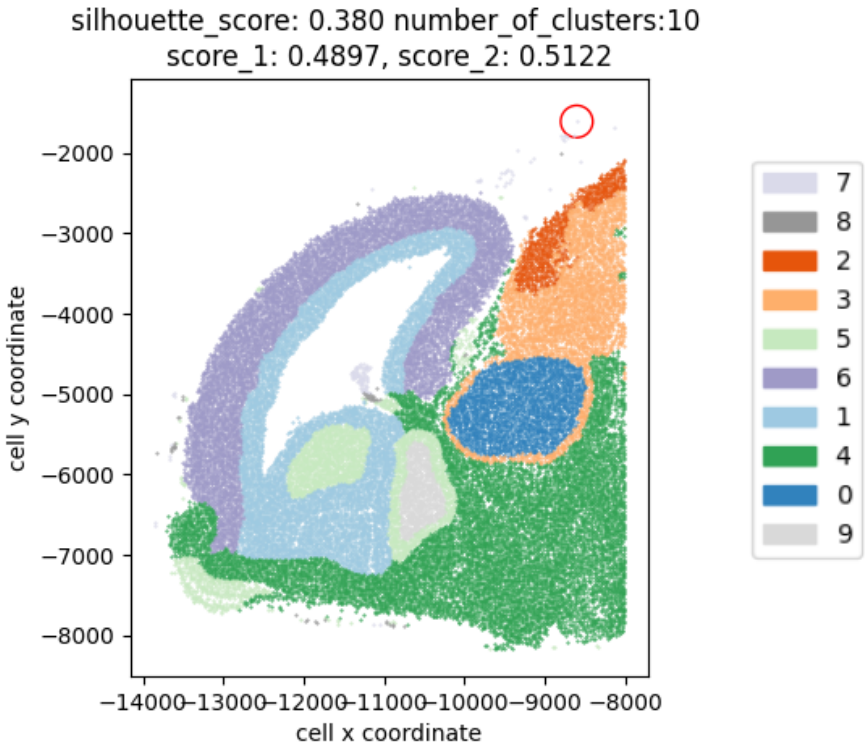


Figure: Clusters

Result 2 - Statistics of clusters

- The mean value for the distance between cells is 1.3, as shown in the image below

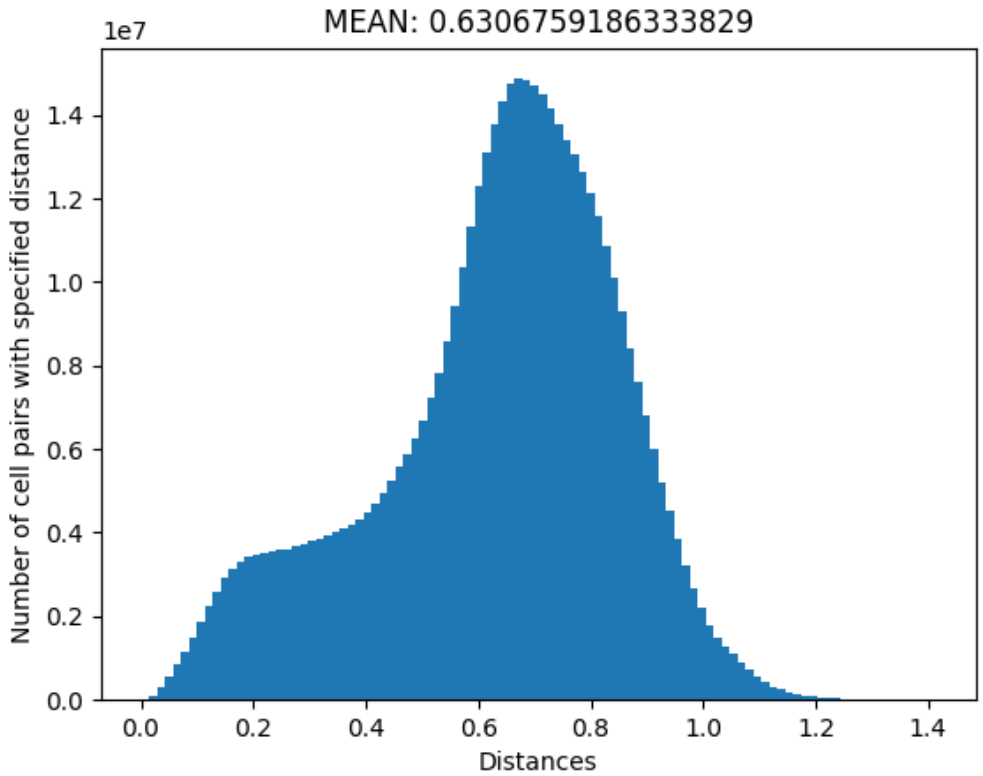


Figure: Distribution of distances between cells for all cells

Result 2 - Homogeneous or Heterogeneous clusters

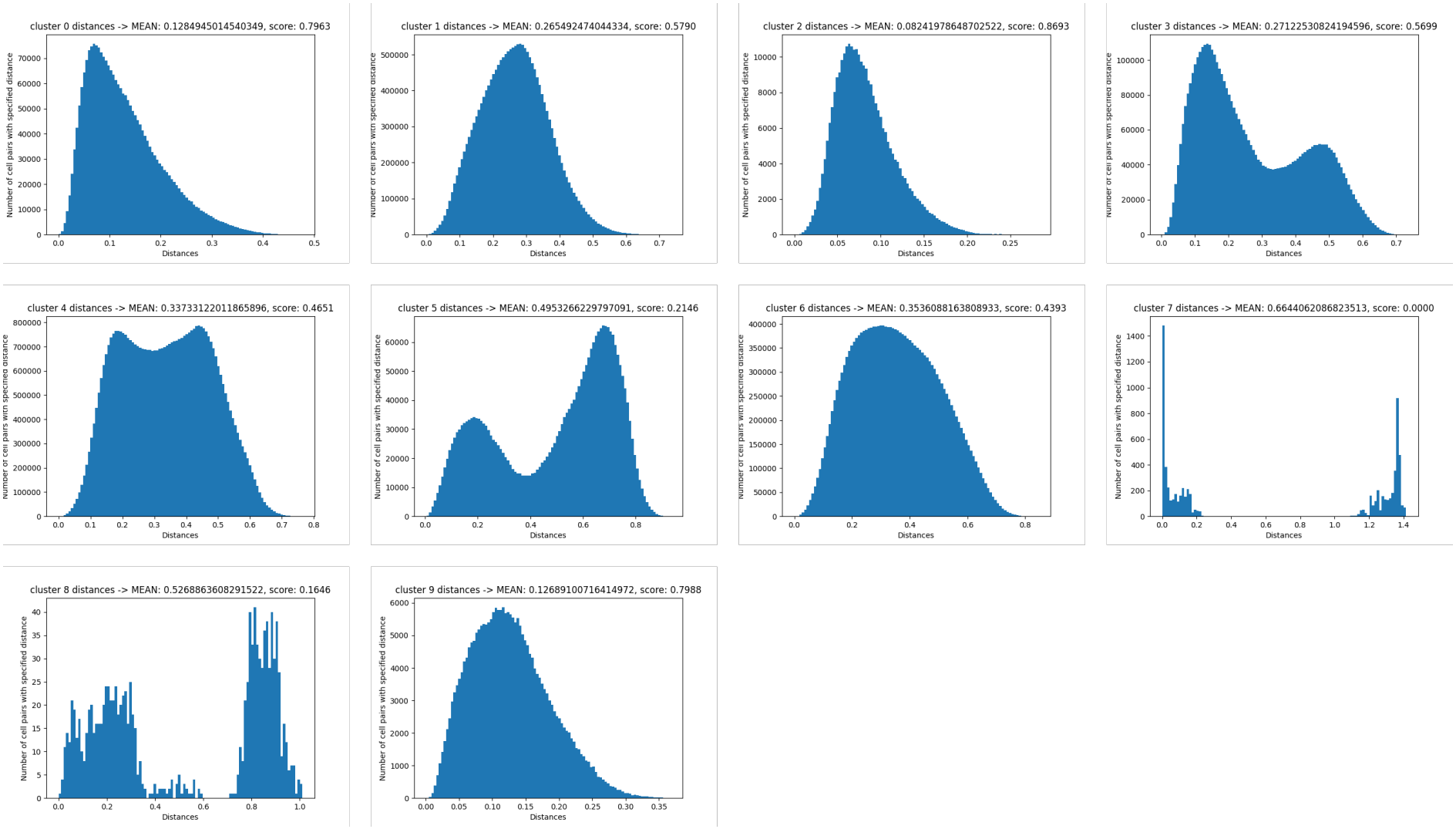


Figure: Distribution of distances between cells for each cluster

Result 2 - Homogeneous or Heterogeneous clusters

- As observed in the image above, there are distributions, particularly for clusters 5, 7, and 8, which may suggest that these clusters need to be divided or that this clustering isn't optimal
- Furthermore, clusters 7 and 8 contain the smallest number of cells, with cluster, having low cell count (shown in the image below)
- Distributions for the other clusters are more promising. Cluster 0, 1, 2, 3 and 9 are homogeneous and clusters 4 and 6 are close to homogeneous

Result 2 - Number of different cell types in each cluster

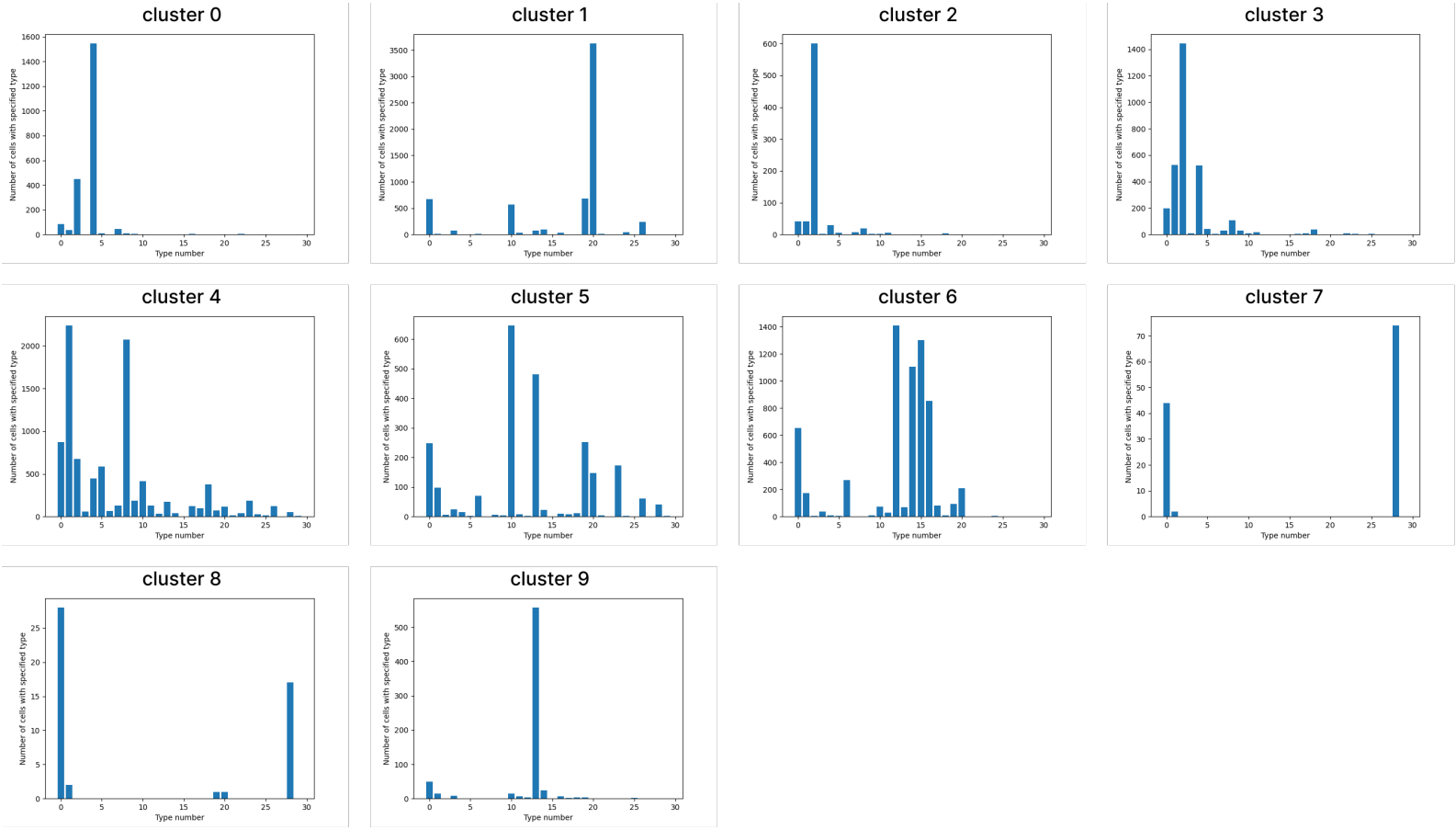


Figure: Number of different cell types in each cluster

Result 2 - Percentage of different cell types in each cluster

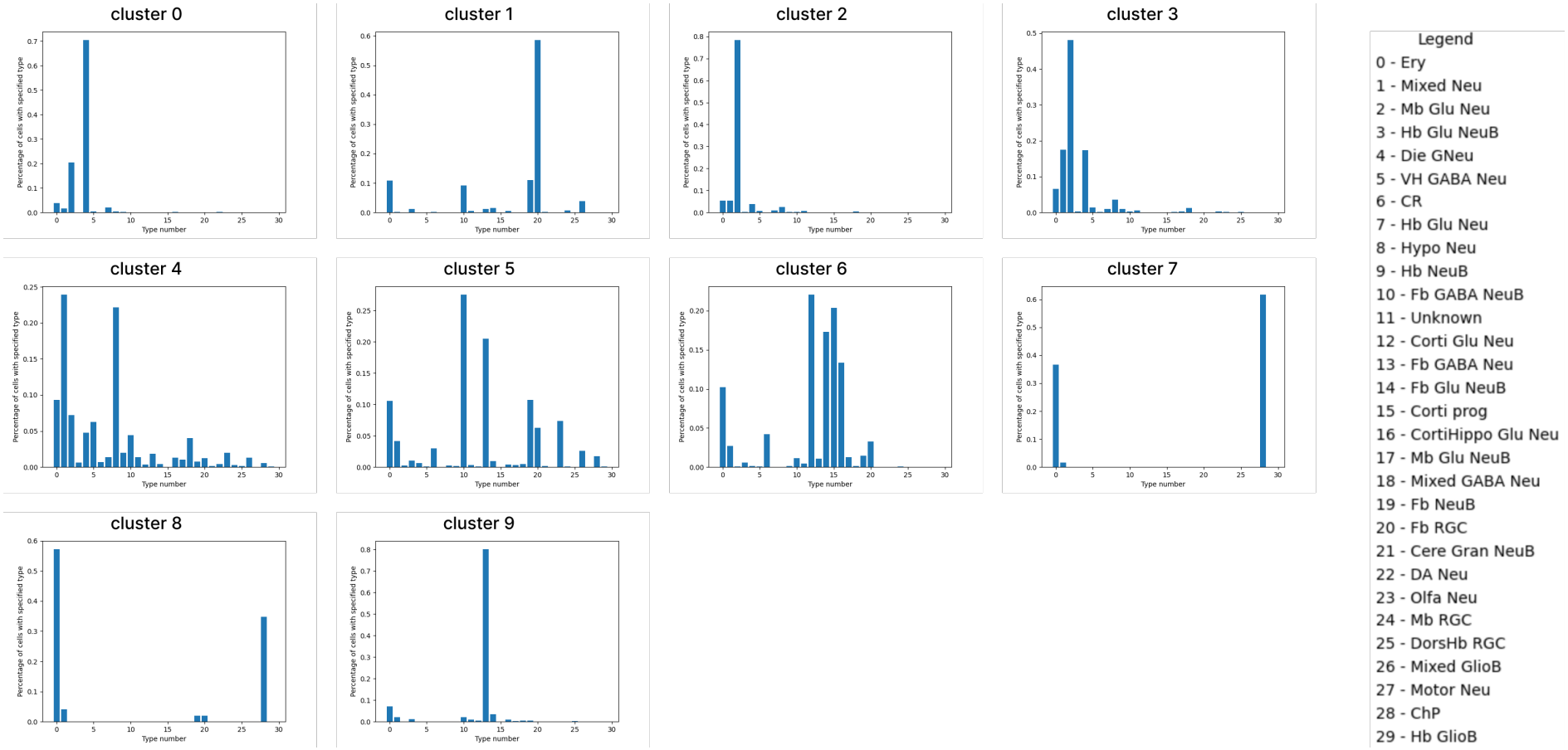


Figure: Percentage of different cell types in each cluster

Result 3

- Results where:
  - `neighborhood_radius = 150` (the radius that corresponds to the red circle in the image below)
  - `distance_function = Hamming` with `hamming_param = 0.3`
  - `linkage_method = average`
  - `threshold = 30`

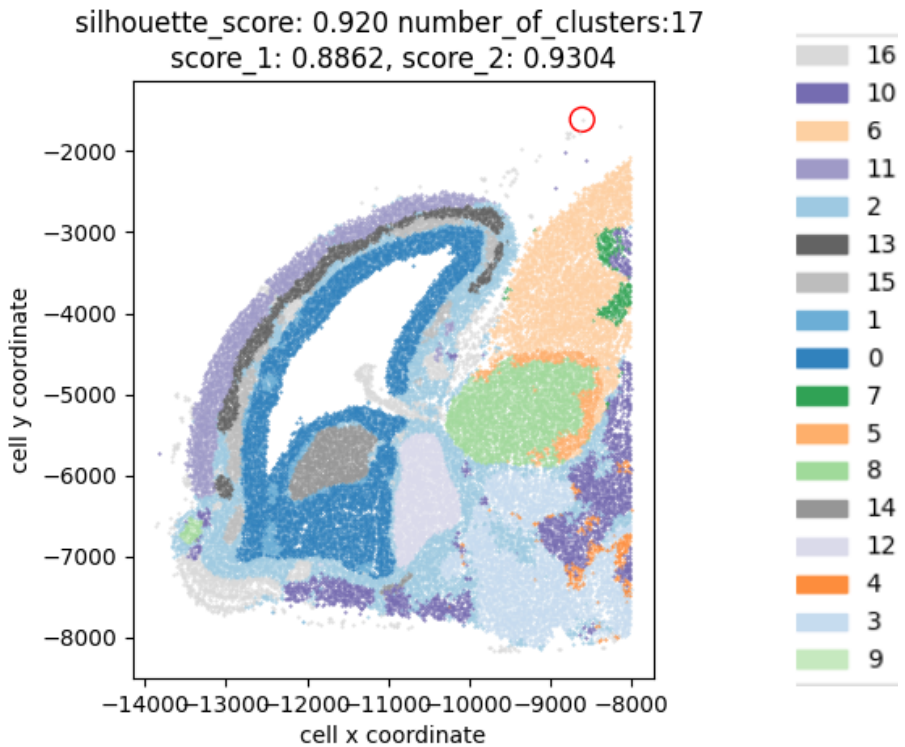


Figure: Clusters

Result 3 - Statistics of clusters

- The maximum value for the distance between cells is 1.3, as shown in the image below

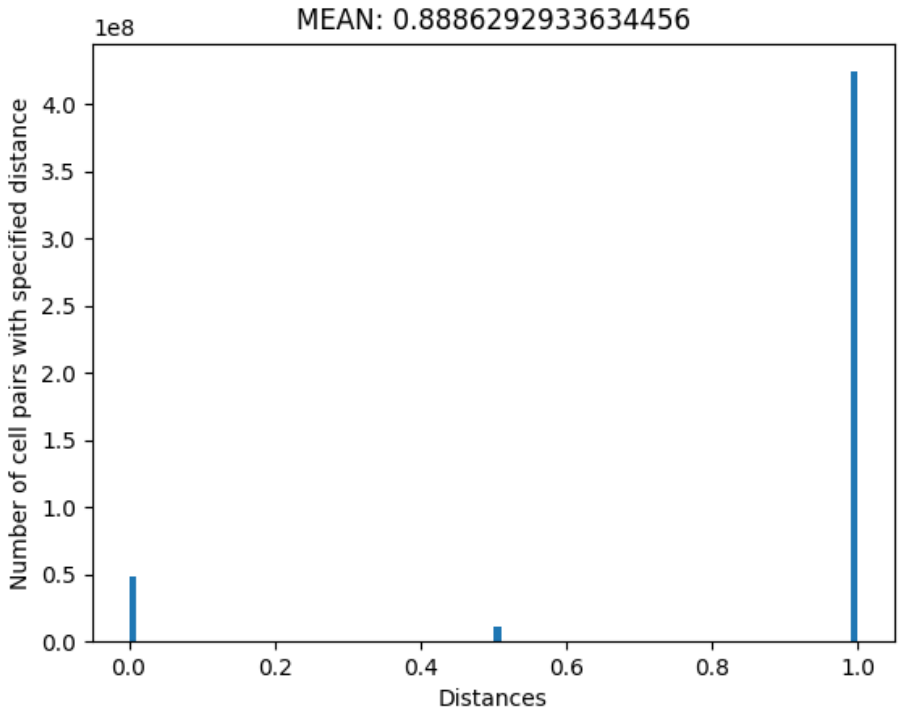


Figure: Distribution of distances between cells for all cells

Result 3 - Homogeneous or Heterogeneous clusters

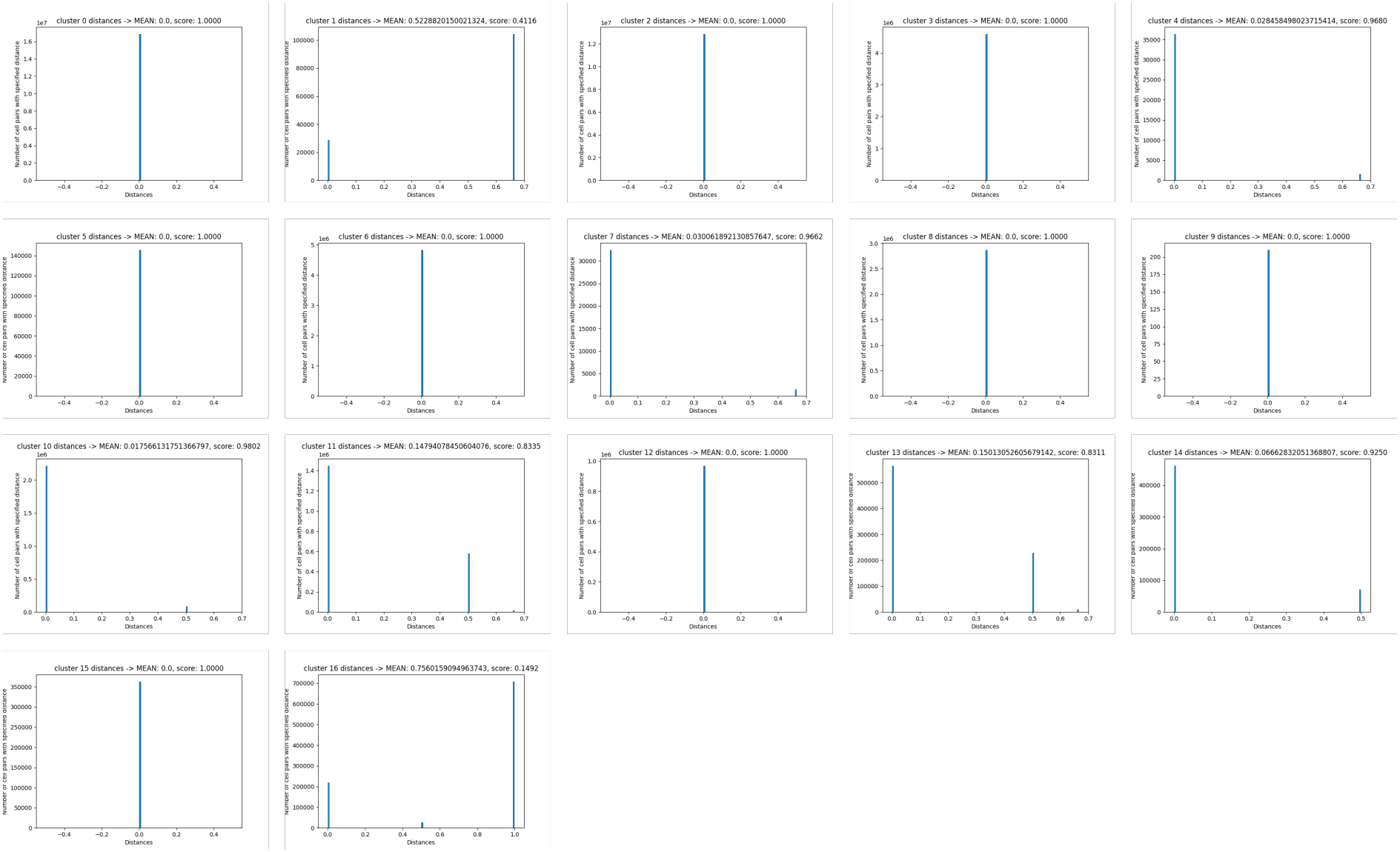
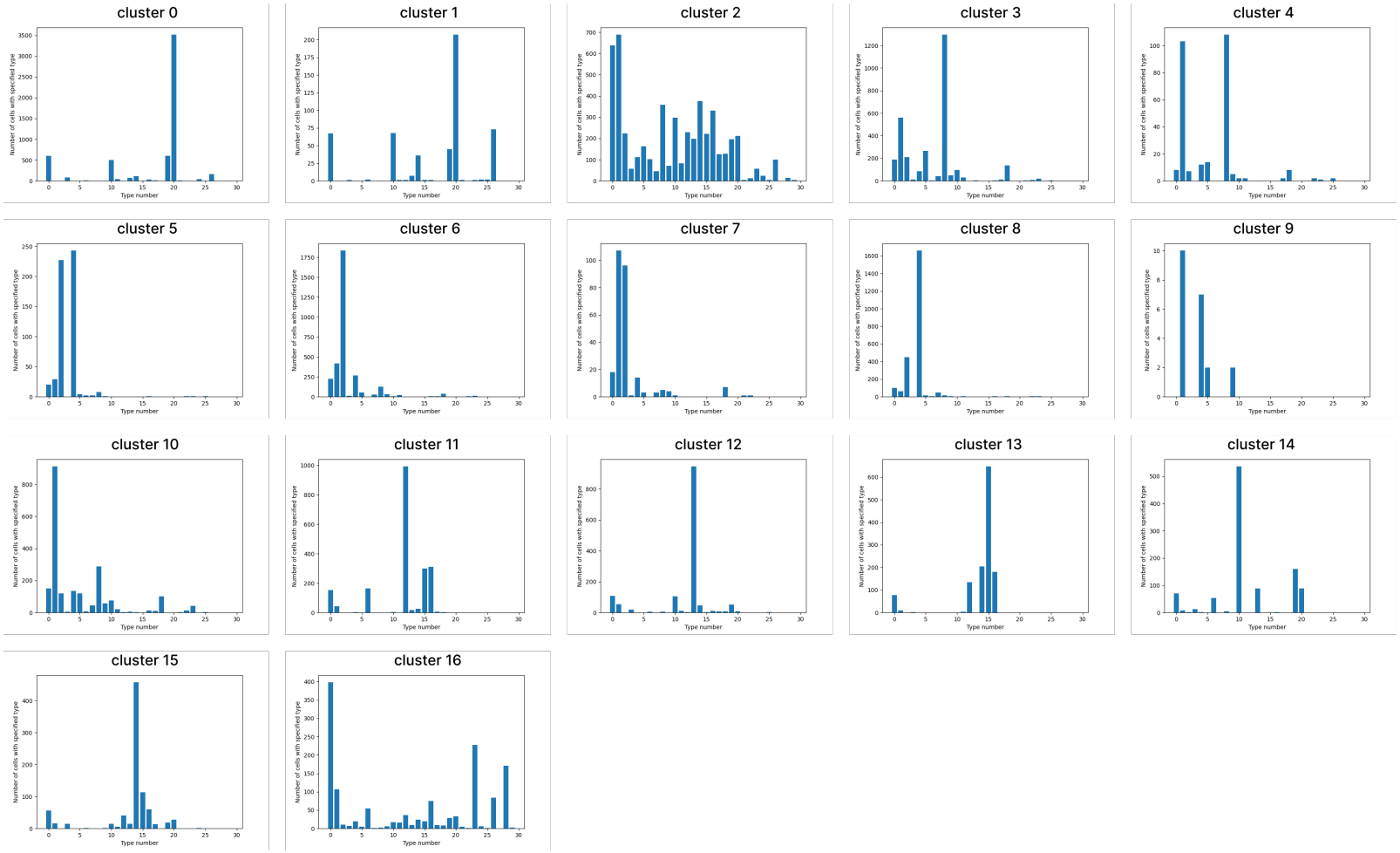


Figure: Distribution of distances between cells for each cluster

Result 3 - Homogeneous or Heterogeneous clusters

- As observed in the image above, there are distributions, particularly for clusters 5, 7, and 8, which may suggest that these clusters need to be divided or that this clustering isn't optimal
- Furthermore, clusters 7 and 8 contain the smallest number of cells, with cluster, having low cell count (shown in the image below)
- Distributions for the other clusters are more promising. Cluster 0, 1, 2, 3 and 9 are homogeneous and clusters 4 and 6 are close to homogeneous

Result 3 - Number of different cell types in each cluster



Legend

0 - Ery
1 - Mixed Neu
2 - Mb Glu Neu
3 - Hb Glu NeuB
4 - Die GNeu
5 - VH GABA Neu
6 - CR
7 - Hb Glu Neu
8 - Hypo Neu
9 - Hb NeuB
10 - Fb GABA NeuB
11 - Unknown
12 - Corti Glu Neu
13 - Fb GABA Neu
14 - Fb Glu NeuB
15 - Corti prog
16 - CortiHippo Glu Neu
17 - Mb Glu NeuB
18 - Mixed GABA Neu
19 - Fb NeuB
20 - Fb RGC
21 - Cere Gran NeuB
22 - DA Neu
23 - Olfa Neu
24 - Mb RGC
25 - DorsHb RGC
26 - Mixed GlioB
27 - Motor Neu
28 - ChP
29 - Hb GlioB

Figure: Number of different cell types in each cluster

Result 3 - Percentage of different cell types in each cluster

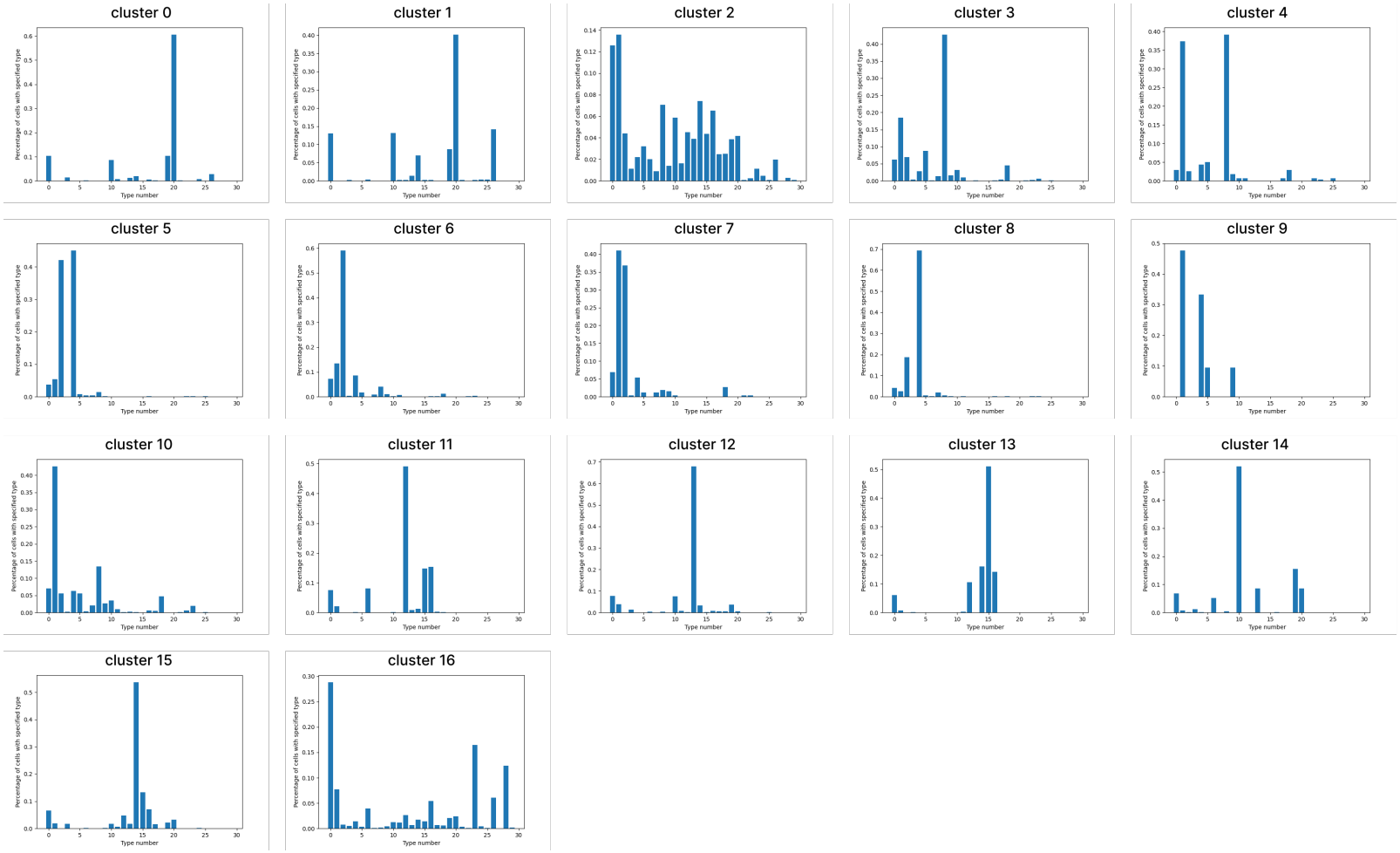


Figure: Percentage of different cell types in each cluster

Result 3 - Increasing the threshold

- Increasing the threshold (from 30 to 45) did result in a lower number of clusters with slightly lower silhouette and total homogeneity score
- Clusters 0 and 1 from the previous clustering has now been merged into single cluster 0. This has resulted in higher homogeneity score for cluster 0 (0.91) than what was observed for clusters 0 (1) and slightly lower score than 1 (0.45) in the previous clustering

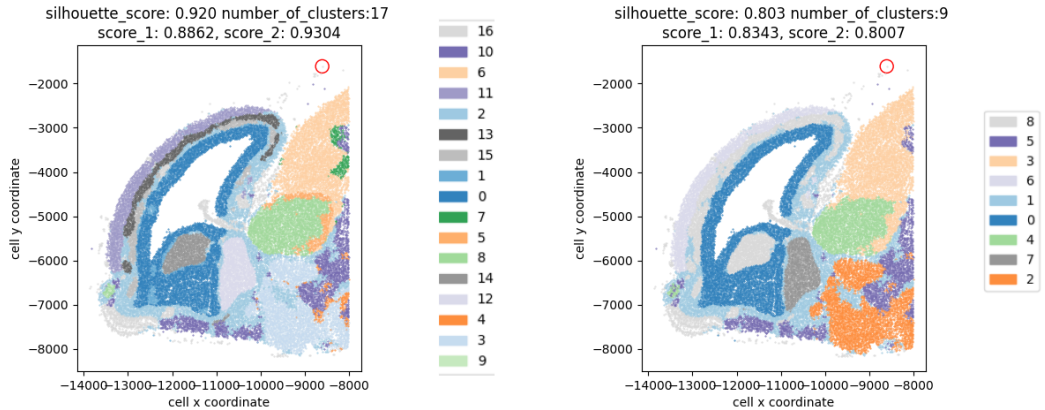


Figure: Clusters with `threshold` 30 and 45, respectively

Result 3 - Increasing the threshold

- Clusters 13 and 14 and 15 from the previous clustering has now been merged into clusters 8. This has resulted in lower homogeneity score for clusters 8 (0.07) than what was observed for the clusters 13 (0.83), 14 (0.92) and 15 (1) in the previous clustering.

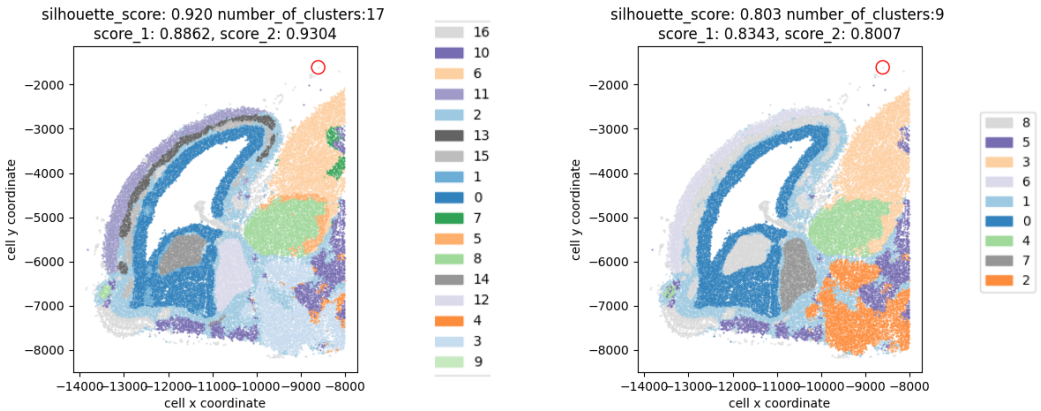


Figure: Clusters with `threshold` 30 and 45, respectively

Result 3 - Increasing the threshold

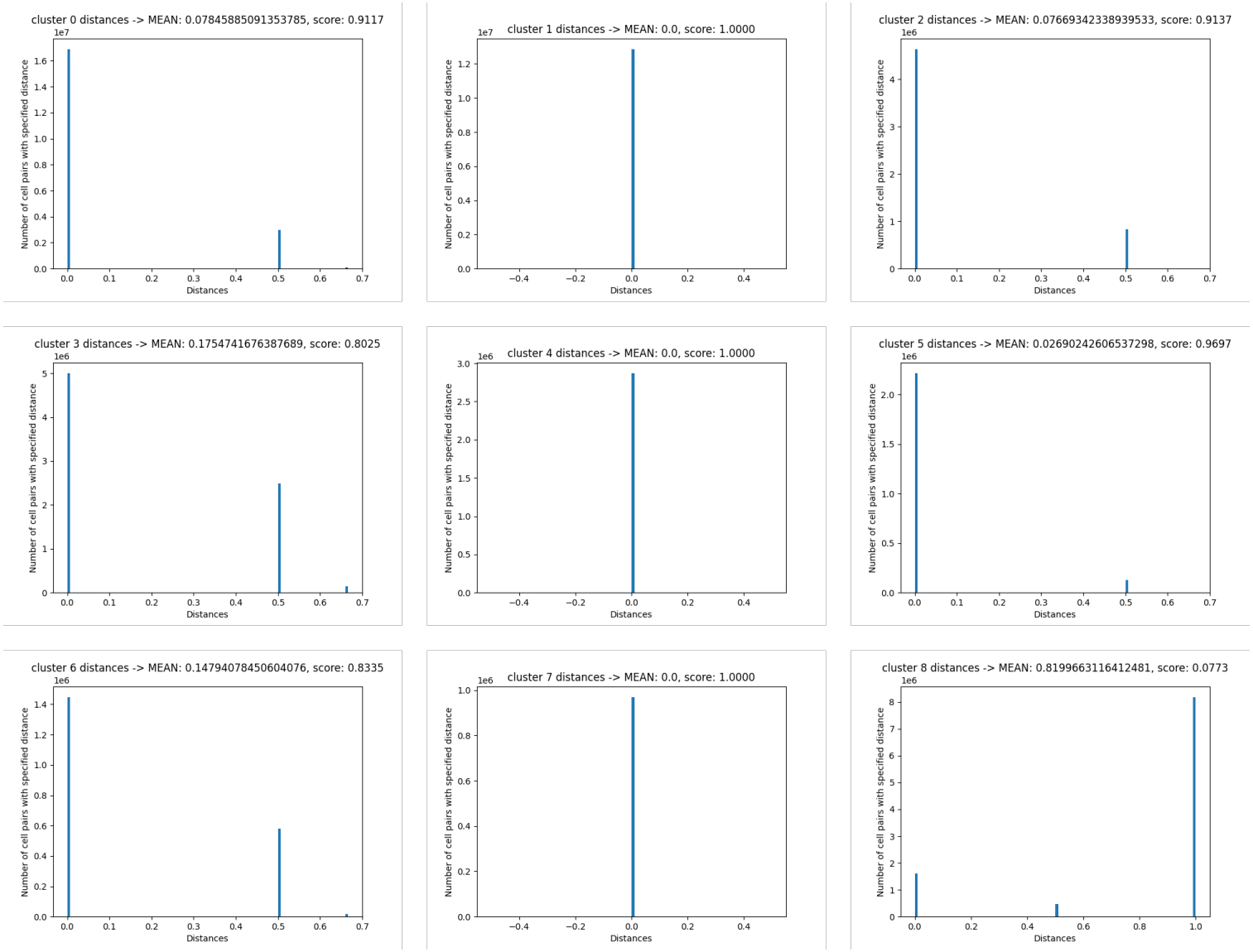


Figure: Distribution of distances between cells for each cluster

Result 3 - Increasing the threshold

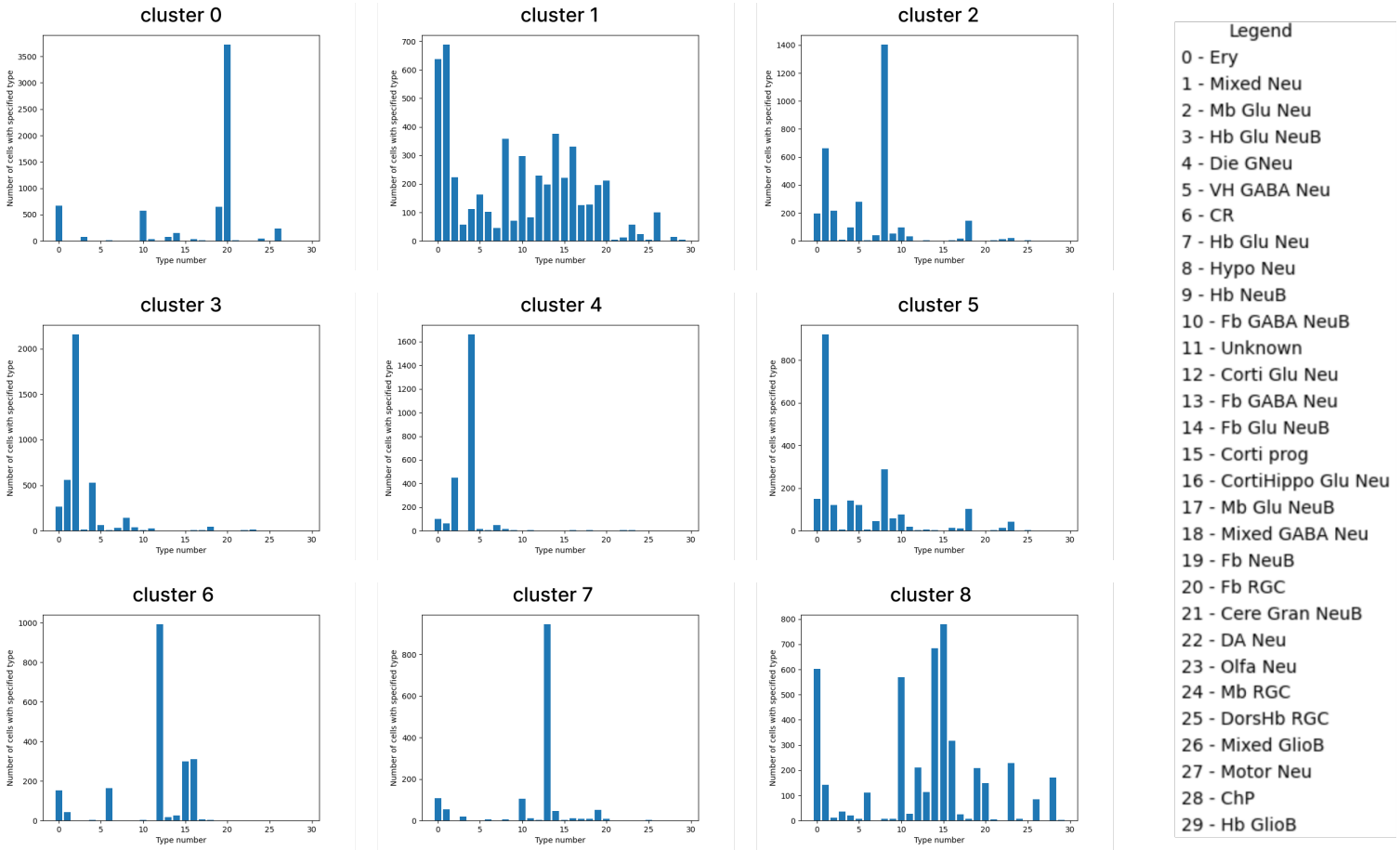


Figure: Number of different cell types in each cluster

Result 3 - Increasing the threshold

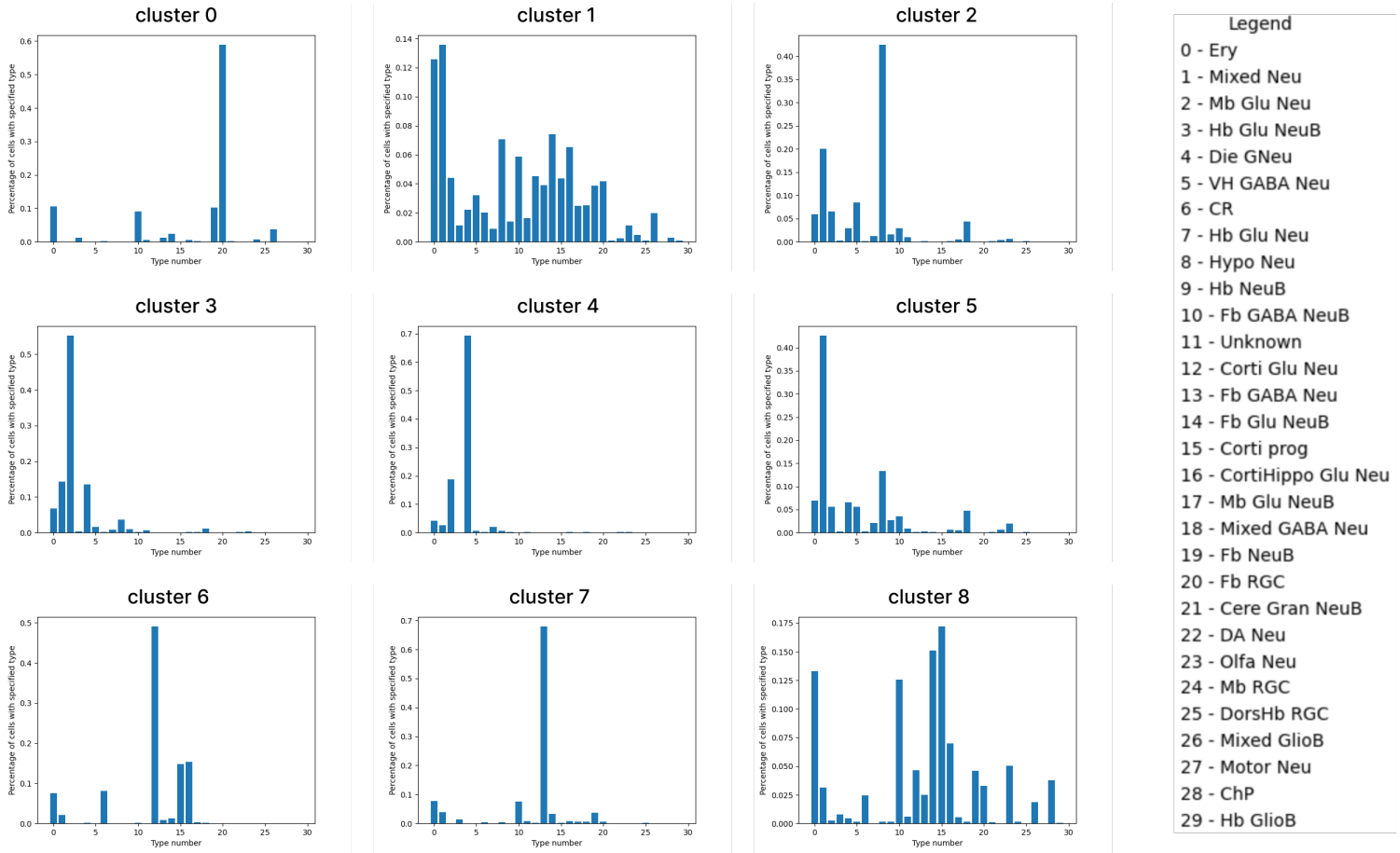


Figure: Percentage of different cell types in each cluster

Conclusion

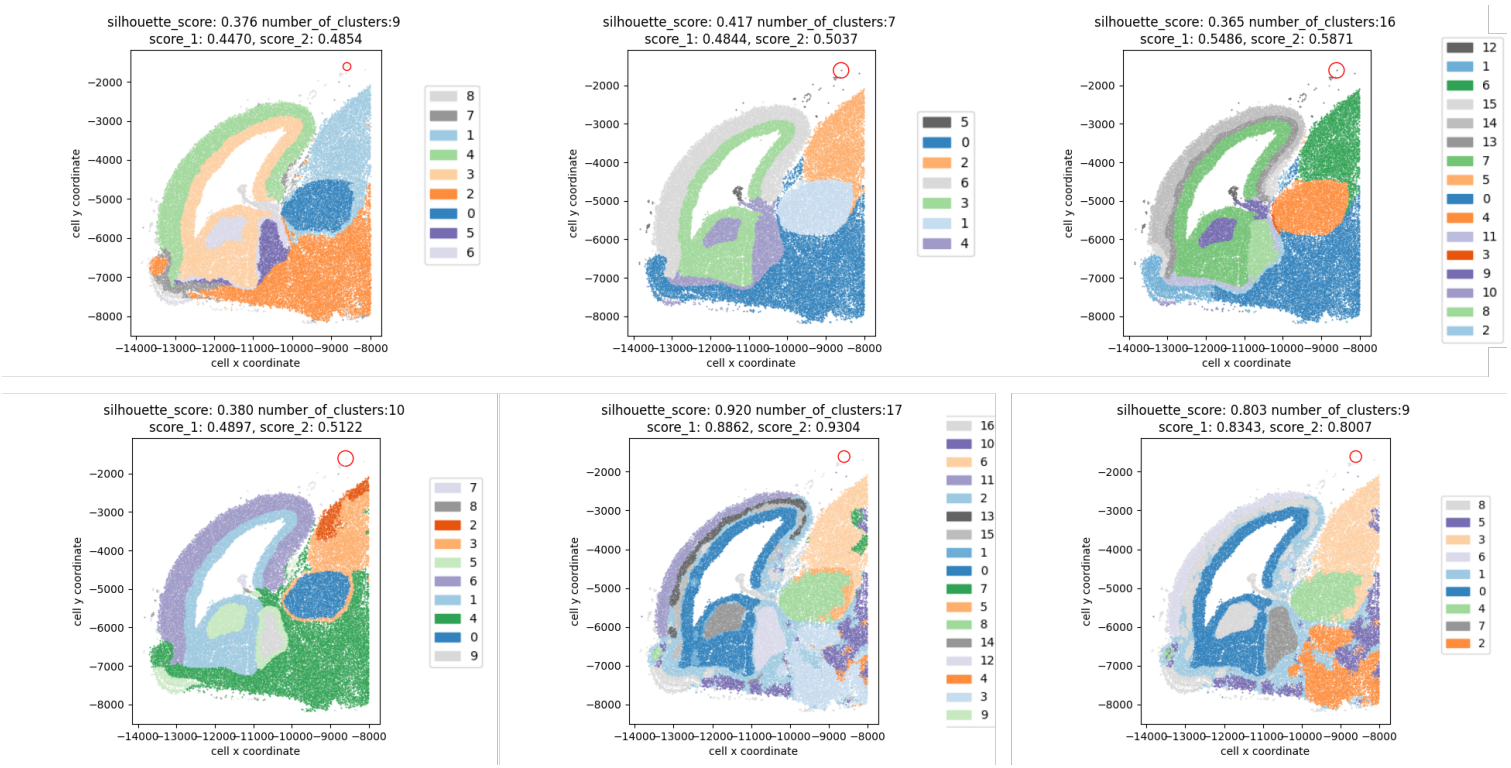


Figure: Clusters for different clusterings - Manhattan (100 200 - different threshold), Euclidean (200) Hamming with parameter 0.3 (150 - different thresholds)

Conclusion

- From the image above, it is evident that several homogeneous clusters are consistently present in almost all of the different clusterings:
  - 0, 1, 4, 0, 8 and 4
  - 5, 8, 9, 12 and 7
  - 3, 3, 7, 1, 0 and 0
  - 1, 2, 6, 3, 6 and 3
- Furthermore, there are clusters that appear as a single cohesive group in one clustering but get divided into smaller, more homogeneous sub-clusters in others