This note explores two measures of clustering stability and their relation to outliers. This leads to insights about the appropriate number of clusters. The Zhengmix4eq and Zhengmix8eq data were studied. For these data sets

* Graphs of the distributions of Misclassification Error (ME) distance [1] conveniently illustrate clustering stability.
* Cell Mean Misclassification Error (MME) characterizes outliers. It is negatively correlated with silhouette scores.
* Removing outliers defined by MME and re-clustering the residual cells yields individual clusterings with less variation (measured by MME) and, in enough cases to justify mention here, consistent consensus clusters. That is, the putative outliers often behave as expected.

Recall the setting

* For each Zhengmix data set, 100 sets of individual clusterings (2-20 clusters) were obtained by spectral clustering of random forest proximities – for both genes and cells.
* Consensus spectral clusterings were calculated from these.
* To evaluate clustering stability, Misclassification Error (ME) distance was calculated for each pair of individual clusterings – for each number of clusters.
* Graphic display of the distribution of ME distance suggests appropriate clusterings. This is analogous to Tibshirani and Walther’s suggestion [2] to choose a number of clusters with high prediction strength.

The balance of this note is devoted to

1. Reviewing the definition of Misclassification Error (ME) distance and illustrating its variation with the number of clusters for the Zhgengmix data sets, for both gene and cell clusters
2. Defining Mean Misclassification Error (MME)
3. Illustrating the negative correlation between MME and silhouette scores
4. Sharing evidence that cells with high MME behave as outliers are expected to
5. Briefly describing the set of python programs in the folder example\_programs\_outlier\_analysis
6. Misclassification Error (ME) distance

For a set **S** of **n** elements and two clusterings of size **k** define the Misclassification Error distance [1] between the clusterings (alternatively called the Minimal Matching distance [3]) by

* Calculating the confusion matrix
* Permuting its rows and columns to maximize the sum of diagonal entries
* Summing the off-diagonal entries and dividing by **n**
* In the following, elements of **S** that contribute to off-diagonal entries are referred to as **misclassified**

This setting, in which multiple pairs of same-size clusterings are compared, is similar to the one in which Tibshirani and Walther define prediction strength [2].

Gene clusters

* The violin plots in Figure 1 show the distribution of ME distance for each number of clusters.
* There is a transition at 6 clusters. The range increases.
  + For smaller clusterings, ME distance rarely exceeds 0.1
  + For larger clusterings the 25th percentile of ME distance exceeds 0.1 in all but two cases:
    - Zhengmix4eq for 8 clusters: the 25th percentile equals 0.098
    - Zhengmix8eq for 7 clusters: the median equals 0.092
* The plots suggest that 5 gene clusters are appropriate for both data sets. In both cases, this is the largest number of clusters with low ME distance.

Cell clusters

* Please refer to the violin plots in Figure 2, which show the distributions of ME distance.
* For each data set there are both decisive limits and ambiguities:
  + Zhengmix4eq: for 5 clusters or fewer, clusterings are very stable; at 6, variability increases. 5-cluster variability is larger than that for 4 clusters.
  + Zhengmix8eq: for 7 clusters or fewer, clusterings are very stable; for 9 or more, variability is much larger. Variability for 8 clusters is in between.
* Conclusion
  + Zhengmix4eq: 4 or 5 cell clusters are plausible
  + Zhengmix8eq: 7 or 8 are plausible, with 7 preferable because its ME distances are smaller

1. Mean Misclassification Error

Informally define Mean Misclassification Error (MME) by example for cell clusterings of Zhengmix data. For a given number of clusters, *MME is defined for each cell*:

* 100 individual clusterings of 3994 cells were obtained, ranging in size from 2 to 20 clusters.
* Fix a number of clusters **k**.
* For each cell, consider all pairs of clusterings
  + Count the number of pairs in which the cell is misclassified.
  + Divide by the total number of pairs (4950).
  + Define the cell’s MME as the quotient.

This parallels the definition in [2] of prediction strength for individual observations. Figure 3 in Tibshirani and Walther’s paper illustrates the connection to outliers.

The distributions of cell MME are illustrated in Figure 3. For the clusterings of interest, this table gives counts by MME range:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Zhengmix4eq** | |  | **Zhengmix8eq** | |
| **MME range** | **4 clusters** | **5 clusters** |  | **7 clusters** | **8 clusters** |
| 0 - 0.01 | 3928 | 3445 |  | 3407 | 1988 |
| 0.01 -0.1 | 24 | 257 |  | 196 | 888 |
| 0.1 – 0.2 | 13 | 110 |  | 81 | 488 |
| 0.2 – 1.0 | 29 | 182 |  | 310 | 630 |

For convenience, say that cells with MME ≤ 0.01 cluster unambiguously.

In each data set, cells in the smaller clustering are more likely to cluster unambiguously, as highlighted.

1. Silhouette scores are negatively correlated with MME

Mean Misclassification Error measures clustering ambiguity for each observation. Since silhouette scores also measure this, it is makes sense that the two quantities are correlated. The correlation is negative because high MME means poor clustering but high silhouette scores mean good clustering.

Figures 4 and 5 contain scatter plots of the data summarized in the following four tables.

Zhengmix4eq

Even though the 4-cluster segmentation has few cells with high MME, this table shows that high MME values are associated with low silhouette scores:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Silhouette score quartile** | | | |  |
| **MME range** | **0** | **1** | **2** | **3** | **Total** |
| 0 - 0.01 | 932 | 999 | 998 | 999 | 3928 |
| 0.01 -0.1 | 24 | 0 | 0 | 0 | 24 |
| 0.1 – 0.2 | 13 | 0 | 0 | 0 | 13 |
| 0.2 – 1.0 | 29 | 0 | 0 | 0 | 29 |
|  |  |  |  |  |  |
| Total | 998 | 999 | 998 | 999 | 3994 |

For 5 clusters, the data are unambiguous:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Silhouette score quartile** | | | |  |
| **MME range** | **0** | **1** | **2** | **3** | **Total** |
| 0 - 0.01 | 616 | 861 | 969 | 999 | 3445 |
| 0.01 -0.1 | 150 | 83 | 24 | 0 | 257 |
| 0.1 – 0.2 | 65 | 40 | 5 | 0 | 110 |
| 0.2 – 1.0 | 167 | 15 | 0 | 0 | 182 |
|  |  |  |  |  |  |
| Total | 998 | 999 | 998 | 999 | 3994 |

Zhengmix8eq

For 7 clusters, this table shows the correlation of MME with silhouette score:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Silhouette score quartile** | | | |  |
| **MME range** | **0** | **1** | **2** | **3** | **Total** |
| 0 - 0.01 | 525 | 897 | 986 | 999 | 3407 |
| 0.01 -0.1 | 128 | 56 | 12 | 0 | 196 |
| 0.1 – 0.2 | 60 | 21 | 0 | 0 | 81 |
| 0.2 – 1.0 | 285 | 25 | 0 | 0 | 310 |
|  |  |  |  |  |  |
| Total | 998 | 999 | 998 | 999 | 3994 |

and for 8 clusters:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Silhouette score quartile** | | | |  |
| **MME range** | **0** | **1** | **2** | **3** | **Total** |
| 0 - 0.01 | 156 | 275 | 617 | 940 | 1988 |
| 0.01 -0.1 | 171 | 390 | 271 | 56 | 888 |
| 0.1 – 0.2 | 170 | 223 | 92 | 3 | 488 |
| 0.2 – 1.0 | 501 | 111 | 18 | 0 | 630 |
|  |  |  |  |  |  |
| Total | 998 | 999 | 998 | 999 | 3994 |

1. Evidence that cells with high Mean Misclassification Error behave like outliers

Although these results are tentative, sharing them may prompt useful follow-up. Two properties are meant by “behave like outliers.” Specifically, if putative outliers are removed and the residual cells are re-clustered

* These “residual clusters” should have fewer outliers. In other words, removing outliers should clean the clustering process, and not make room for other cells to take their place.
* The residual clusters should agree with the original clusters – restricted, of course, to the residual cells.

Two sets of results are summarized here.

* In the first, clustering uses the random forest proximities obtained for all 3994 cells, restricting to residuals. This gives surprisingly good results, perhaps because of the connection between spectral clusters and graphs.
* In the second, the residual cells were clustered from first principles – starting with UMI counts. Results are not as sharp.

Analysis using previously calculated proximities

For each of the four clusterings explored here in detail (4 and 5 clusters for Zhengmix4eq, 7 and 8 for Zhengmix8eq) these steps were performed

* Define cells as outliers if MME ≥ 0.1
* Remove them from the random forest proximity arrays
* Re-cluster the residual cells; call these “residual clusterings” as opposed to “original clusterings”
* Calculate MME for the residual cells.
* Compare each residual consensus clustering with the original clustering of the same size, restricting to residual cells. Count the number of misclassified cells.

In all four cases

* As tabulated on the next page, the residual consensus clusterings (of the appropriate size) agree exactly with the original clusters, restricted to the residual cells.
* The distribution of MME for the residual clusterings is lower than for the original clusterings. Informally: after outliers are removed, no large number of residual cells become outliers.

The differences between distributions of MME for the original and residual clusterings are illustrated in violin plots in Figure 6.

Figure 7 contains scatter plots of MME vs silhouette scores for residual cells, for comparison with Figures 4 and 5.

For the four data sets, this table reports the number of misclassified residual cells, for 2-10 consensus clusters. The top data row lists the number of residual cells:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Zhengmix4eq** | | **Zhengmix8eq** | |
|  | **4 clusters** | **5 clusters** | **7 clusters** | **8 clusters** |
| **Residual cells:** | **3952** | **3702** | **3603** | **2877** |
| Clusters: 2 | 1 | 2 | 1 | 403 |
| 3 | 0 | 3 | 6 | 200 |
| 4 | **0** | 1 | 17 | 7 |
| 5 | 6 | **0** | 15 | 770 |
| 6 | 38 | 81 | 10 | 31 |
| 7 | 49 | 174 | **0** | 0 |
| 8 | 49 | 197 | 203 | **0** |
| 9 | 54 | 212 | 379 | 41 |
| 10 | 127 | 166 | 409 | 275 |

Read the right column as

* Deleting 1117 outliers leaves 2877 residual cells
* After re-clustering, the 7 and 8 consensus clusterings agree exactly with the original clusters (for the 2877 residual cells)

First principle re-clustering of residual cells

The Zhengmix8eq data set was selected for this analysis, because its 8 cluster segmentation has 1117 cells with MME ≥ 0.1, the largest number of outliers of the four clusterings studied here in depth. Bluntly, it is the worst of the four. This leaves 2877 residual cells for re-clustering.

The following steps were performed:

* UMI counts were selected for the 2877 cells.
* Binomial deviance was calculated; 170 genes were selected per the randomization criteria.
* Standardized null residuals were calculated for the 170 genes with highest binomial deviance.
* Synthetic data sets were created, random forest classification and clustering were performed for genes.
* Gene Misclassification Error distance is illustrated in the upper chart in Figure 8. It suggests that 7 gene clusters are appropriate
* Gene means were calculated for the 7 clusters.
* Synthetic data sets were created, random forest classification and clustering were performed for cells. Cell Misclassification Error distance is illustrated in the lower chart in Figure 8. Distributions are tight for several segmentations with 12 or fewer clusters:
  + All of their distribution medians are less than 0.1
  + The maxima for the 4, 6, and 7-clusterings are less than 0.08
* MME was calculated for all cells. Distributions are tabulated on the next page for 6-9 clusters. These bracket 7 and 8, which seemed most plausible based on results previously discussed. For comparison, the corresponding distributions for all 3994 cells are also tabulated.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Zhengmix8eq data – 2877 residual cells** | | | |
| **MME range** | **6 clusters** | **7 clusters** | **8 clusters** | **9 clusters** |
| 0 - 0.01 | 2598 | 2638 | 1829 | 1666 |
| 0.01 -0.1 | 96 | 84 | 468 | 533 |
| 0.1 – 0.2 | 47 | 32 | 182 | 86 |
| 0.2 – 1.0 | 136 | 123 | 398 | 592 |
|  |  |  |  |  |
| % unambiguous ( ≤ 0.01) | 90.3 | 91.7 | 63.5 | 57.9 |
| % outliers ( ≥ 0.1) | 6.4 | 5.4 | 20.2 | 23.6 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Zhengmix8eq data – all 3994 cells** | | | |
| **MME range** | **6 clusters** | **7 clusters** | **8 clusters** | **9 clusters** |
| 0 - 0.01 | 3509 | 3407 | 1988 | 2091 |
| 0.01 -0.1 | 206 | 196 | 888 | 669 |
| 0.1 – 0.2 | 59 | 81 | 488 | 218 |
| 0.2 – 1.0 | 220 | 310 | 630 | 1016 |
|  |  |  |  |  |
| % unambiguous ( ≤ 0.01) | 87.9 | 85.3 | 49.8 | 52.4 |
| % outliers ( ≥0.1) | 7.0 | 9.8 | 28.0 | 30.9 |

Although the differences between the two tables are not dramatic, they move in the desired direction. For the residual cells

* The percentage unambiguously clustered is higher.
* Outlier percentages are lower.

Among the residual clusterings, the one with 7 clusters has

* the highest percentage of cells that cluster unambiguously and
* the lowest percentage of outliers
* Cell consensus clusters for the 2877 residual cells were compared to the original clusters. The segmentation with 7 clusters has the smallest number of misclassified cells, 113, for a ME distance of 0.04

|  |  |  |
| --- | --- | --- |
| Clusters | **misclassified** | **ME distance** |
| 2 | 575 | 0.20 |
| 3 | 234 | 0.08 |
| 4 | 859 | 0.30 |
| 5 | 194 | 0.07 |
| 6 | 209 | 0.07 |
| **7** | **113** | **0.04** |
| 8 | 151 | 0.05 |
| 9 | 467 | 0.16 |
| 10 | 568 | 0.20 |

1. Program folder example\_programs\_outlier\_analysis

The folder contains 10 python programs for outlier analysis. They include data preparation, analysis, and graphic output.

As written, they perform a complete analysis for the 8-cluster segmentation of the Zhengmix8eq data set. The necessary input files could be prepared by the suite in the repository folder example\_programs, appropriately modified to analyze the Zhengmix8eq data set (instead of Zhengmix4eq, as posted).

1. cell\_silhouette\_scores\_from\_mean\_proximities.py
2. plot\_MME\_vs\_cell\_silhouette\_scores\_from\_mean\_proximities.py
3. identify\_cell\_cluster\_outliers.py
4. proximities\_cells\_drop\_outliers.py
5. cell\_clusters\_\_individual\_and\_consensus\_drop\_outliers.py
6. analyze\_cell\_clustering\_stability\_drop\_outliers.py
7. cell\_mean\_proximities\_drop\_outliers.py
8. cell\_silhouette\_scores\_from\_mean\_proximities\_drop\_outliers.py
9. plot\_MME\_vs\_cell\_silhouette\_scores\_from\_mean\_proximities\_drop\_outliers.py
10. extract\_counts\_class\_labels\_for\_cell\_subset.py

Descriptions

1. cell\_silhouette\_scores\_from\_mean\_proximities.py

* Calculate silhouette scores from the consensus clusterings and the mean proximity array.

1. plot\_MME\_vs\_cell\_silhouette\_scores\_from\_mean\_proximities.py

* Create scatter plots, as in Figure 4, from silhouette scores and cell Mean Misclassification Error.

1. identify\_cell\_cluster\_outliers.py

* For a specified number of clusters, identify all cells with MME ≥ 0.1

1. proximities\_cells\_drop\_outliers.py

* For each random forest proximity array, exclude rows and columns corresponding to cell outliers identified in step 1.

1. cell\_mean\_proximities\_drop\_outliers.py

* Deal with the mean proximity array in the same way as the individual random forest proximities in step 4.
* That is, exclude rows and columns corresponding to cell outliers identified in step 3.

1. cell\_clusters\_\_individual\_and\_consensus\_drop\_outliers.py

* Calculate individual and consensus clusterings from the proximities prepared in step 4.
* This program, with the ones listed below, except the very last – which extracts cell UMI counts and class labels for non-outliers – are redundant. They duplicate functionality of codes that are already posted, that could easily be modified. They are included for convenience, to make it easier to perform similar calculations on data sets with different file names in different folders.

1. analyze\_cell\_clustering\_stability\_drop\_outliers.py

* Calculate ARI and ME distance for cell clusters calculated in step 6.
* Calculate MME for cells.

1. cell\_silhouette\_scores\_from\_mean\_proximities\_drop\_outliers.py

* Perform the calculations of step 1 – with the residual mean proximity array calculated in step 5 and the consensus clusters calculated in step 6.

1. plot\_MME\_vs\_cell\_silhouette\_scores\_from\_mean\_proximities\_drop\_outliers.py

* Generate scatter plots as in step 2, for residual cells only
* Input data are MME calculated in step 7 and silhouette scores calculated in step 8.

1. extract\_counts\_class\_labels\_for\_cell\_subset.py

* Prepare data to perform a “first principle re-clustering of residual cells”
* Before running this program, it is necessary to create the folder in which CSV output files will be written. For example, in the code provided
  + the input folder is D:\scRNA-seq\Zhengmix8eq
  + and output is D:\scRNA-seq\Zhengmix8eq\_2877

Figure 1

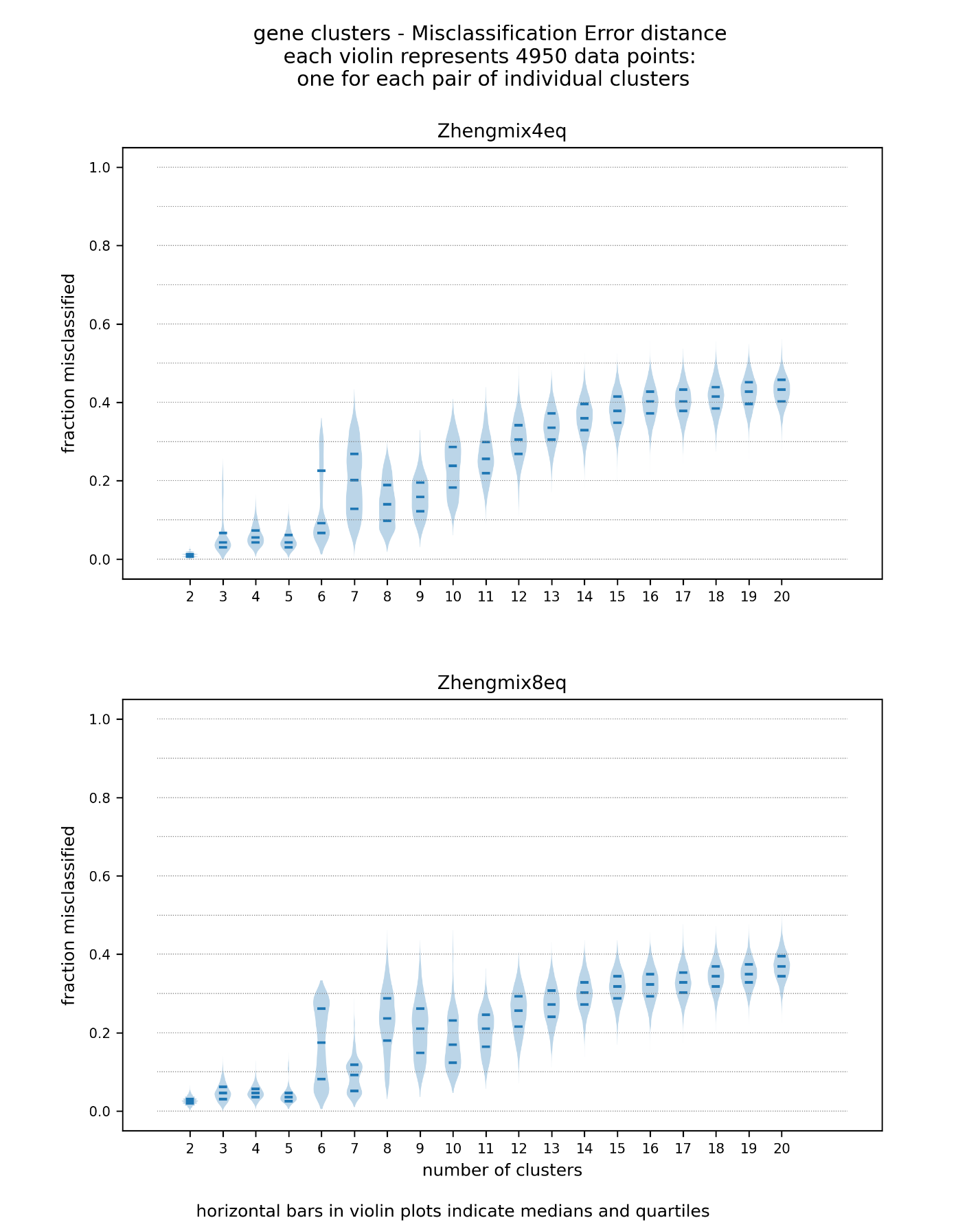


Figure 2

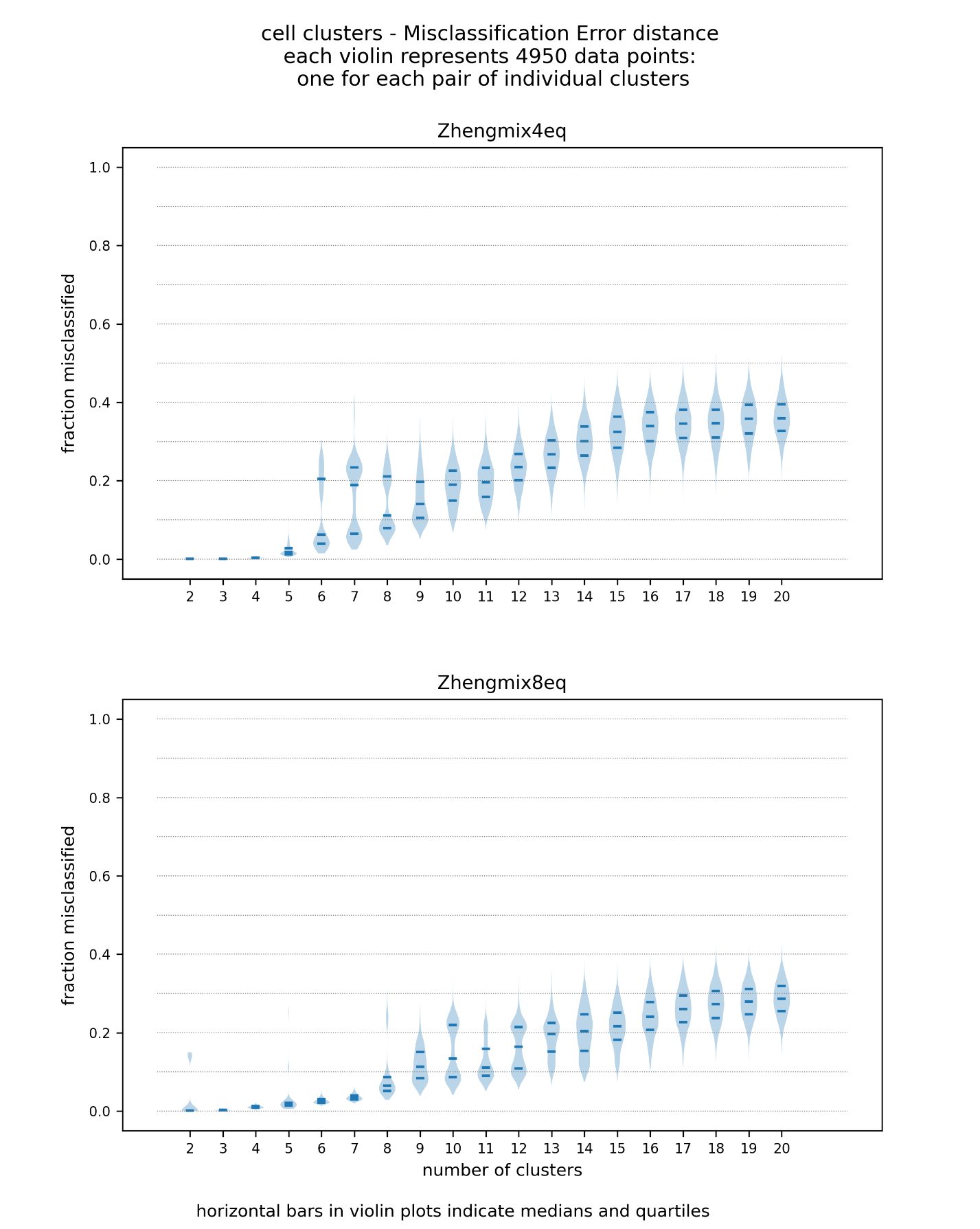


Figure 3

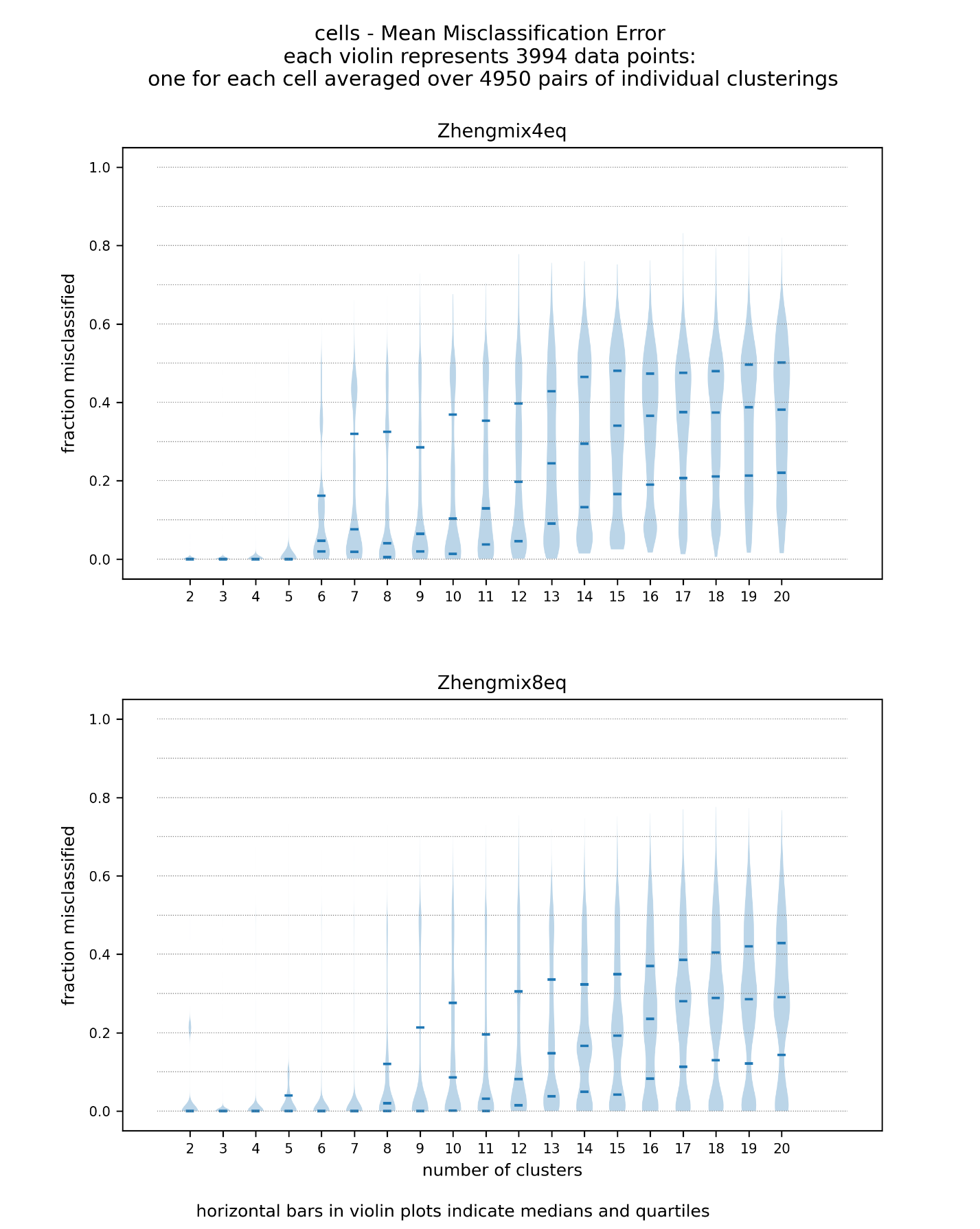


Figure 4

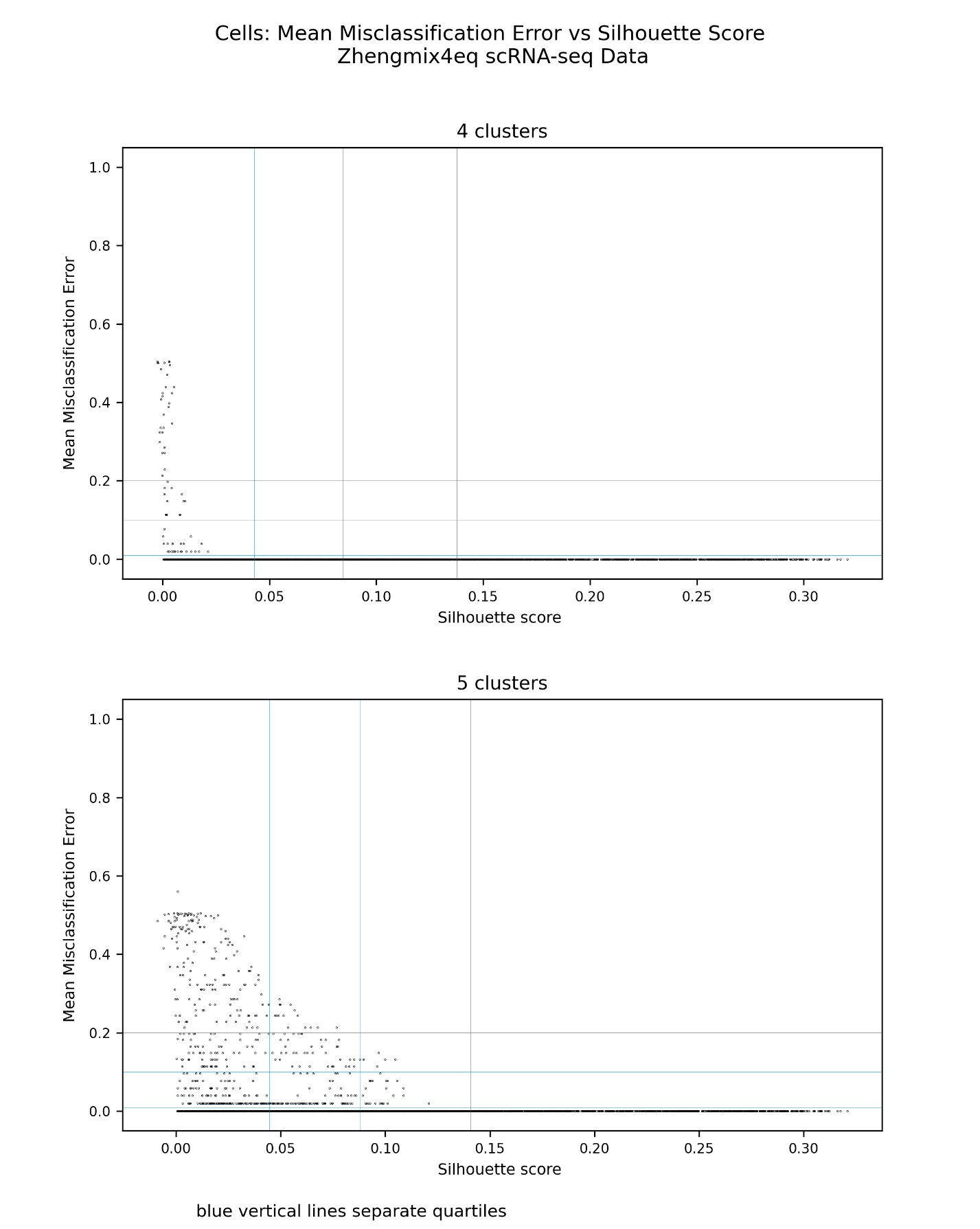


Figure 5

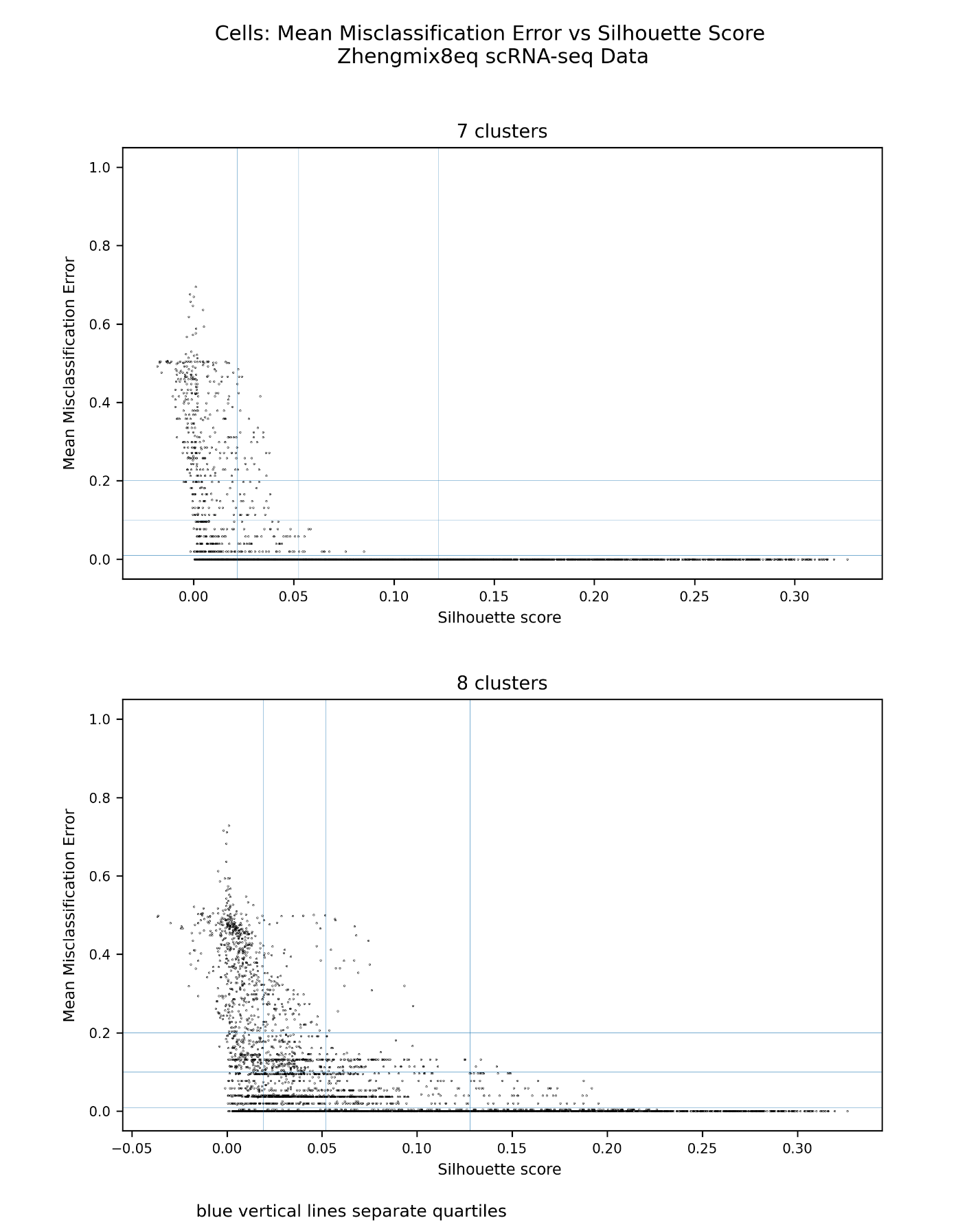


Figure 6

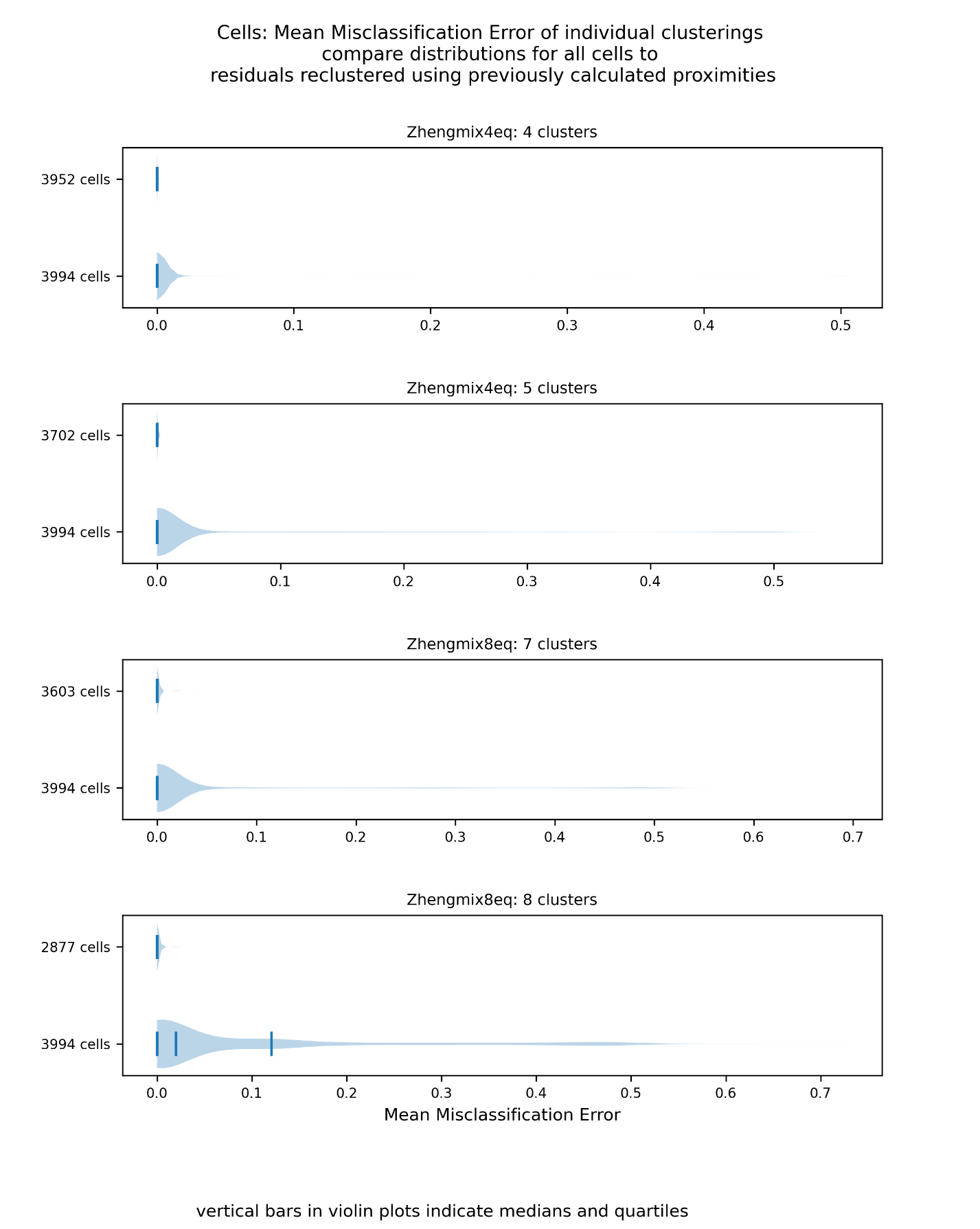


Figure 7

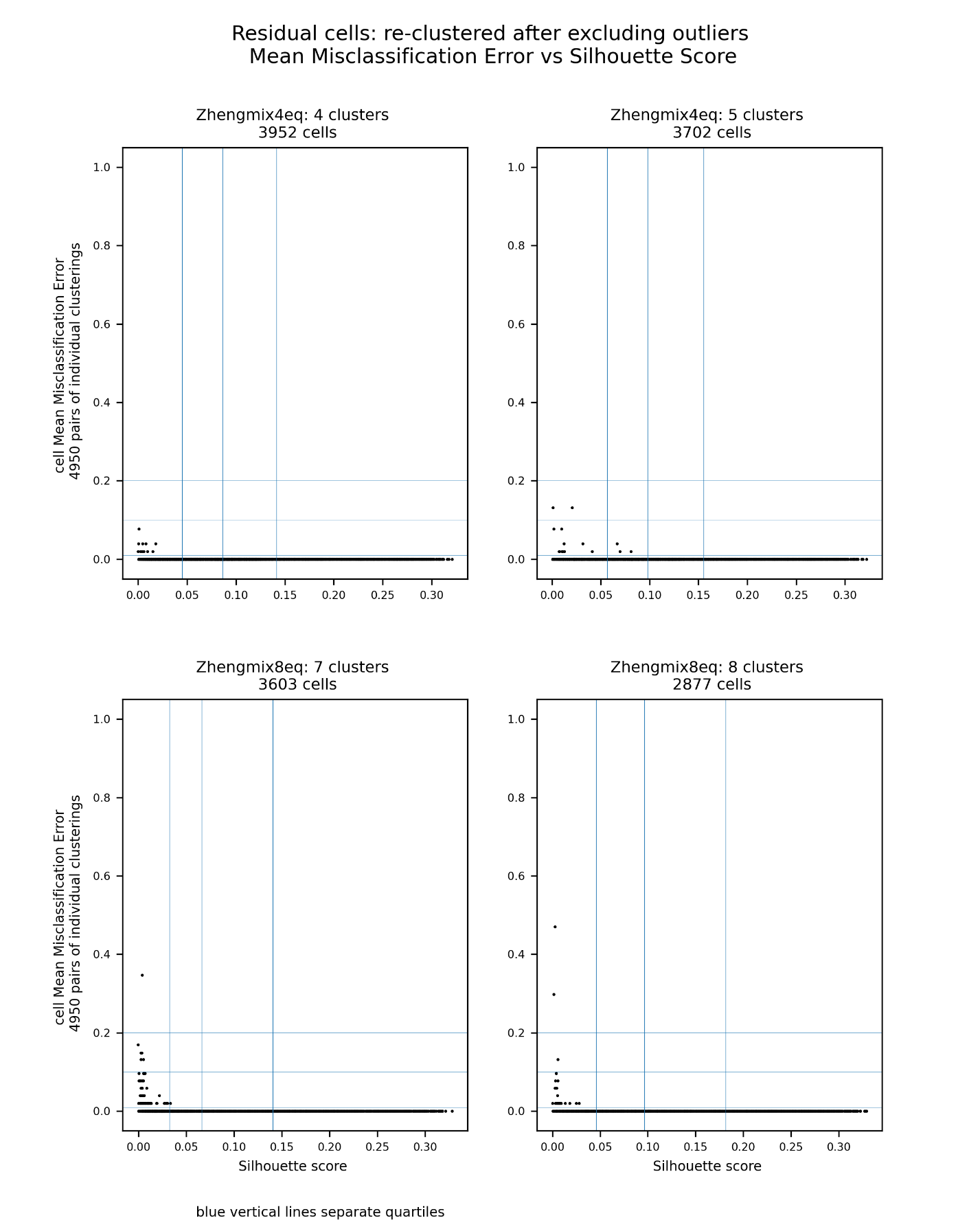
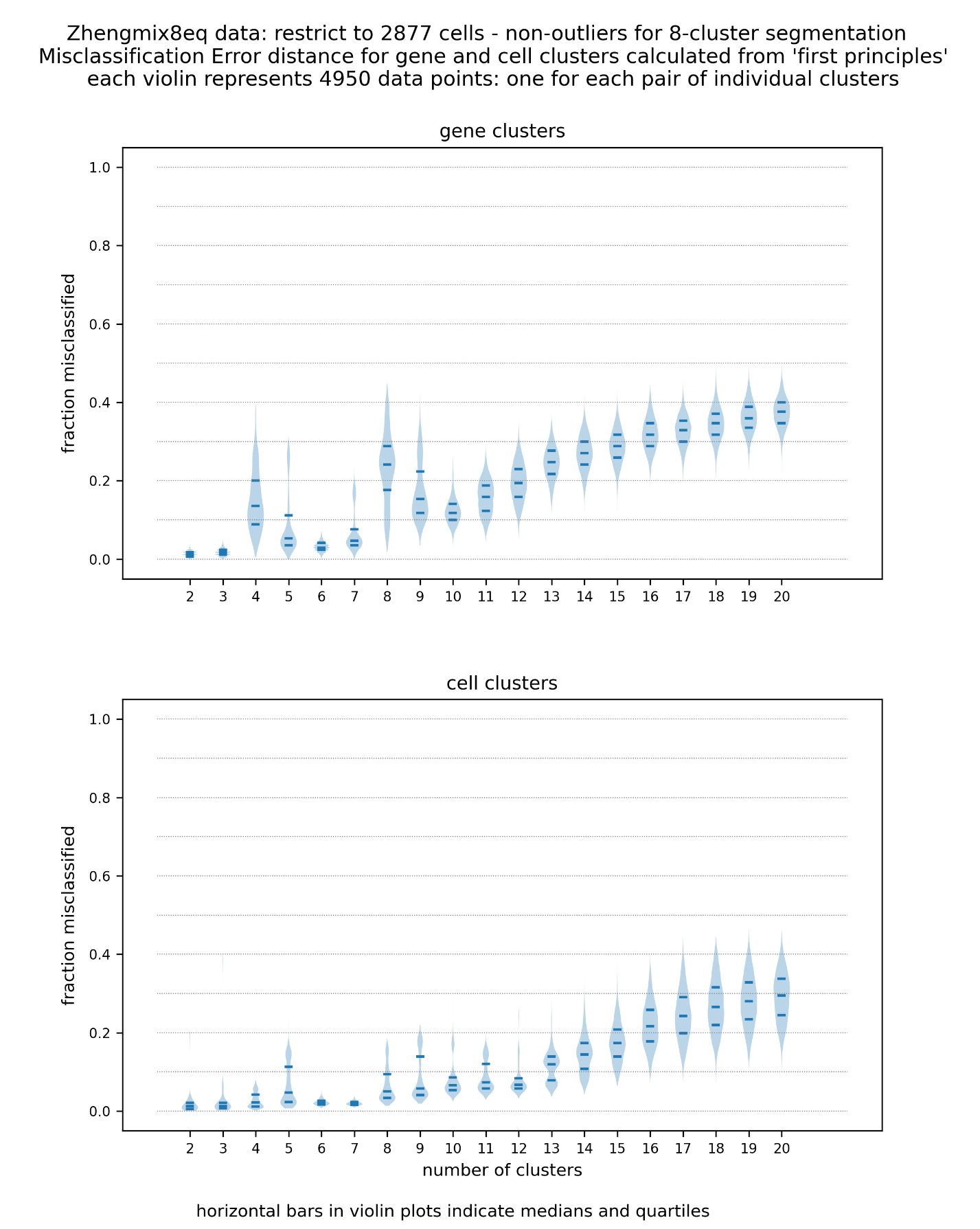


Figure 8



**References**

1. Meila M: The stability of a good clustering. <https://stat.washington.edu/sites/default/files/files/reports/2014/tr624.pdf>
2. Tibshirani R, Walther G: Cluster validation by prediction strength, *Journal of Computational and Graphical Statistic* 2005, 14:3, 511-528, DOI: 10.1198/106186005X59243
3. Ben-David S, Luxburg U: Relating Clustering Stability to Properties of Cluster Boundaries. Proceedings of the 21st Annual Conference on Learning Theory (COLT 2008), 379-390