This set of programs clusters single cell RNA-Seq data, represented by UMI counts, with no metric assumptions and no logarithmic transformation to treat zero counts.

The approach is unconventional, using

* binomial deviance [1] to rank genes for filtering
* random forest classification [2,3,4] to produce proximities to cluster genes and cells
* spectral consensus clustering using the random forest proximities [5]
* distributions of adjusted Rand index or Misclassification Error distance [6] to determine an appropriate number of clusters
* Laplacian scores [7] computed with random forest cell proximities to confirm the gene ranks and provide additional validation of the methodology.

This summary has 4 sections:

1. Outline the approach and highlight the novel analytic features
2. Discuss methods to evaluate cluster stability
3. Share results for the Zhengmix4eq and Zhengmix8eq data sets [8]
4. List of programs that perform complete analysis for the Zhengmix4eq data set

**1. Outline of approach**

Filter genes using binomial deviance

Genes are ranked by binomial deviance, following Townes et al. [1]. This suggests a principled method to select genes using randomization:

* Compute binomial deviance for all genes.
* Compute binomial deviance for randomized data – since randomized data are biologically meaningless, these values should lower-bound the appropriate threshold:
  + Permute each cell’s data – this corrupts gene data.
  + Compute the binomial deviance for the randomized data set.
  + Use the maximum binomial deviance for the randomized data as a filtering threshold.
  + Alternatively, perform multiple permutations, calculate binomial deviance for each permuted data set, and use the median of the maxima.
* There is no *a-priori* guarantee that this approach will yield **any** genes for analysis. As an example, for the ERCC data set studied by Townes et al.
  + The maximum binomial deviance equals 1,927.
  + For permuted data, the **minimum** exceeds 15,000.
  + **No genes** satisfy the proposed filtering criterion.
  + This is appropriate: Townes et al. write that “We refer to this dataset as the technical replicates negative control as there is no biological variability whatsoever and, in principle, each expression proﬁle should be the same.”

Having selected a gene set, the null residuals are calculated as in [1] and the data are standardized for each gene (mean 0, standard deviation 1).

Cluster genes; reduce dimensionality; cluster cells

Clustering is a two-stage process

* Cluster genes, choose an appropriate number of clusters, use these to reduce data dimensionality
* Cluster cells

Gene clustering follows the approach sketched by Breiman and Cutler [2], and elaborated by Shi and Horvath [3]. The insight that classification methods can be used for clustering is recognized in [4]:

* Randomize the data to be clustered to create a synthetic data set.
* Use a random forest classifier to discriminate genuine from randomized data.
* This yields a proximity matrix for the genuine data.
* Perform spectral clustering with the proximities.
* Repeat this process this multiple times, for distinct synthetic data sets, then compute consensus clusterings.

Evaluate cluster stability; choose an appropriate number of gene clusters.

The gene clusterings are used to reduce dimensions

* For each gene cluster, calculate each cell’s mean deviance residual.
* This low-dimensional representation is used to cluster cells.
* As with PCA, this uses linear combinations of expression data to define the low-dimension coordinates.

To cluster cells, follow the same approach as for genes: perform spectral consensus clustering of random forest proximities, evaluate cluster stability, and choose an appropriate number of cell clusters.

The example programs produce two heat maps: one represents the standardized data; the second scales each gene’s residuals as percentiles. For presentation, clusters are ordered with an algorithm proposed by Ding [9] to order leaves of a dendrogram.

The cell proximities are used to calculate Laplacian scores [7] for genes. For the Zhengmix data sets analyzed with this approach, Laplacian scores are strongly negatively correlated with binomial deviance.   This is further evidence for deviance being an appropriate filtering criterion.

The programs in this posting include calculations of the correlation of Laplacian scores with binomial deviance and a scatter plot.

**2. Evaluating cluster stability**

The calculation of multiple individual clusterings, then consensus clusterings derived from these, naturally lends itself to evaluating stability:

* The adjusted Rand index and Misclassification Error distance [6] can be computed for each number of clusters for
  + all pairs of individual clusterings
  + each individual clustering compared to the consensus clusters
* The gap statistic [10] is calculated with distances derived from random forest proximities.
* Misclassification Error distances are also calculated for each clustered observation (gene or cell). Observations that resist clustering (outliers) may indicate biological or data preparation issues – or flaws in this analytical approach. Exploratory analyses (not shared here) suggest that a small number of observations account for most misclassification errors for clusterings of potential interest.
* Comparing multiple sets of clusterings is in the spirit of Dudoit and Fridlyand’s Clest [11] and Tibshirani and Walther’s prediction strength [12]. Calculating Misclassification Error for each observation parallels the definition of prediction strength for individual observations.

In addition to comparing individual clusterings, independent sets of consensus clusterings may be compared:

* For the Zhengmix data, 100 sets of synthetic data were generated for random forest classifications.
* This produced 100 sets of clusterings
* In addition to consensus clusters based on the 100 individual clusterings
  + The 100 sets were split into two groups, consensus clusterings were computed for each set of 50 individual clusterings
  + The 100 sets into four groups, consensus clusterings were computed for each set of 50 individual clusterings
* Misclassification Error was computed for
  + The 2 sets of consensus clusters each derived from 50 individual clusterings
  + The 6 pairs of consensus clusterings each derived from 25 individual clusterings.

Although more work needs to be done, the variation of the distributions of Misclassification Error or ARI with the number of clusters suggests a defensible approach to choosing the number of clusters, analagous to Clest and prediction strength. For Misclassificaton Error:

* For each number of clusters K other than the largest (in the examples, clusterings with 2-20 clusters were cosidered) compare the distribution of its Misclassification Error with all larger clusterings – i.e. all clusterings with L>K clusters
* If the distributions for all of the larger clusterings are larger (measure to be determined, Kolmogorov-Smirnov statistic, perhaps), choose K clusters.

Of course, Misclassification Error distances for consensus clusters can contribute to the choice of number of clusters; this too requires further work.

**3. Results for Zhengmix data sets**

In the following, Misclassification Error distances are presented for both individual and consensus clusterings.

* For individual clusterings, results are summarized in violin plots.
* Multiple consensus clusterings were derived from disjoint subsets of individual clusterings
  + Two sets of consensus clusterings each derived from 50 individual clusterings
  + Four sets consensus clusterings each derived from 25 individual clusterings

These allow comparisons between

* + one pair of clusterings (each based on 50 individual clusterings) for each number of clusters and
  + six pairs of clusterings (each based on 25 individual clusterings) – combinations of the 4 sets

respectively.

Zhengmix4eq

* 164 genes were selected by the binomial deviance / randomization algorithm
* 5 **gene clusters** are suggested by adjusted Rand index, Misclassification Error, and gap statistic.

Figure 1 shows violin plots of the distributions of

* + Misclassification Error for the 4950 pairs of individual clusterings, for 2-20 clusters. Each violin represents 4950 data points.
  + Mean Misclassification Error for each gene. Each violin represents 164 data points – one for each gene. Each gene’s data is its mean Misclassification Error, averaged over the 4950 pairs of individual clusterings.

For the consensus clusters derived from sets of 50 and 25 individual clusterings

|  |  |
| --- | --- |
| # sets of clusters | Misclassification Error distance |
| 2 | 0.012 (2 of 164 genes misclassified) |
| 4 | 0.000 0.006 0.006 0.006 0.012 0.012 |

* 4 or 5 **cell clusters** are suggested by adjusted Rand index and Misclassification Error; the gap statistic finds no satisfactory choice in range 2-19 clusters.

Figure 2 shows violin plots of the distributions of

* + Misclassification Error for the 4950 pairs of individual clusterings, for 2-20 clusters. Each violin represents 4950 data points.
  + Mean Misclassification Error for each cell. Each violin represents 3994 data points – one for each cell. Each cell’s data is its mean Misclassification Error, averaged over the 4950 pairs of individual clusterings.

For the 4-cluster cell consensus clusterings error distances are very small

|  |  |
| --- | --- |
| # sets of clusters | Misclassification Error distance |
| 2 | 0.00075 (3 of 3994 cells misclassified) |
| 4 | 0.0005 0.00075 0.00075 0.00075 0.0010 0.0013 |

For the 5-cluster cell consensus clusterings error distances are larger

|  |  |
| --- | --- |
| # sets of clusters | Misclassification Error distance |
| 2 | 0.0038 (15 of 3994 cells misclassified) |
| 4 | 0.0033 0.0040 0.0043 0.0043 0.0063 0.0065 |

* Laplacian scores: correlation with binomial deviance equals - 0.71
* Comparing the 4 consensus clusters derived from all 100 individual clusters with the ground truth cell classes provided with the data:

Misclassification Error distance equals 0.014 (54 cells misclassified)

Zhengmix8eq

* 195 genes are selected by binomial deviance / randomization algorithm
* 5 **gene clusters** are suggested by adjusted Rand index and Misclassification Error; 7 are suggested by the gap statistic

Figure 3 shows violin plots of the distributions of

* + Misclassification Error for the 4950 pairs of individual clusterings, for 2-20 clusters.
  + Mean Misclassification Error for each gene.

Consensus clustering results for both are consistent:

5 gene clusters

|  |  |
| --- | --- |
| # sets of clusters | Misclassification Error distance |
| 2 | 0.010 (2 of 195 genes misclassified) |
| 4 | 0.005 0.010 0.015 0.015 0.021 0.021 |

7 gene clusters

|  |  |
| --- | --- |
| # sets of clusters | Misclassification Error distance |
| 2 | 0.005 (1 of 195 genes misclassified) |
| 4 | 0.010 0.010 0.021 0.056 0.067 0.067 |

* For both the 5 and 7 gene cluster options, **7 cell clusters** are suggested by adjusted Rand index and Misclassification Error; gap statistic finds no satisfactory choice in range 2-19 clusters

For the cell clusterings based on 5 gene clusters, Figure 4 shows violin plots of the distributions of

* + Misclassification Error for the 4950 pairs of individual clusterings
  + Mean Misclassification Error for each cell.

Consensus clustering results for both 7-cluster segmentations are consistent:

5 gene clusters, 7 cell clusters:

|  |  |
| --- | --- |
| # sets of clusters | Misclassification Error distance |
| 2 | 0.0068 (27 of 3994 cells misclassified) |
| 4 | 0.0058 0.0060 0.0063 0.0068 0.0073 0.0088 |

7 gene clusters, 7 cell clusters:

|  |  |
| --- | --- |
| # sets of clusters | Misclassification Error distance |
| 2 | 0.0073 (29 of 3994 cells misclassified) |
| 4 | 0.0055 0.0063 0.0070 0.0075 0.0083 0.0108 |

Note: these two cell clusterings are consistent: the Misclassification Error distance equals 0.07: 280 cells are misclassified.

* Laplacian scores: correlation with binomial deviance equals - 0.62
* Comparing the 8-cluster consensus clustering with the 8 ground truth cell classes gives much larger Misclassification Errors than the Zhengmix4eq data:

|  |  |
| --- | --- |
| # gene clusters | Misclassification Error distance |
| 5 | 0.25 (979 of 3994 cells misclassified) |
| 7 | 0.27 (1068 cells misclassified) |

**4. Example programs for the Zhengmix4eq data set**

All programs are in python except the R code to extract data from the DuoClustering2018 package.

These 17 programs perform a complete analysis for the Zhengmix4eq data set [8].

1. extract\_data\_from\_package\_DuoClustering2018.r
2. compute\_binomial\_deviance.py
3. data\_prep\_for\_RF\_\_genes.py
4. RandomForestClassifier\_for\_proximity\_by\_gene.py
5. gene\_clusters\_\_individual\_and\_consensus.py
6. analyze\_gene\_clustering\_stability.py
7. calculate\_mean\_proximity\_matrix\_\_genes.py
8. gap\_statistic\_\_from\_mean\_proximity\_matrix\_\_genes.py
9. data\_prep\_for\_RF\_\_cells.py
10. RandomForestClassifier\_for\_proximity\_by\_cell.py
11. cell\_clusters\_\_individual\_and\_consensus.py
12. analyze\_cell\_clustering\_stability.py
13. calculate\_mean\_proximity\_matrix\_\_cells.py
14. gap\_statistic\_\_from\_mean\_proximity\_matrix\_\_cells.py
15. data\_prep\_and\_heat\_maps.py
16. Laplacian\_scores\_for\_genes.py
17. compare\_cell\_clusters\_with\_ground\_truth.py

For the Zhengmix data sets discussed here these programs took approximately 7 hours, running Anaconda/Intel python on a Dell Mobile Precision workstation with Intel Core Xeon E-2176M processor. The purpose of this posting is to share the approach; if it is useful, the combination of choosing parameters for efficient processing (for example, 48 synthetic data sets instead of 100) and advice from programmers who understand python efficiency should contribute to performance improvements.

Figure 1

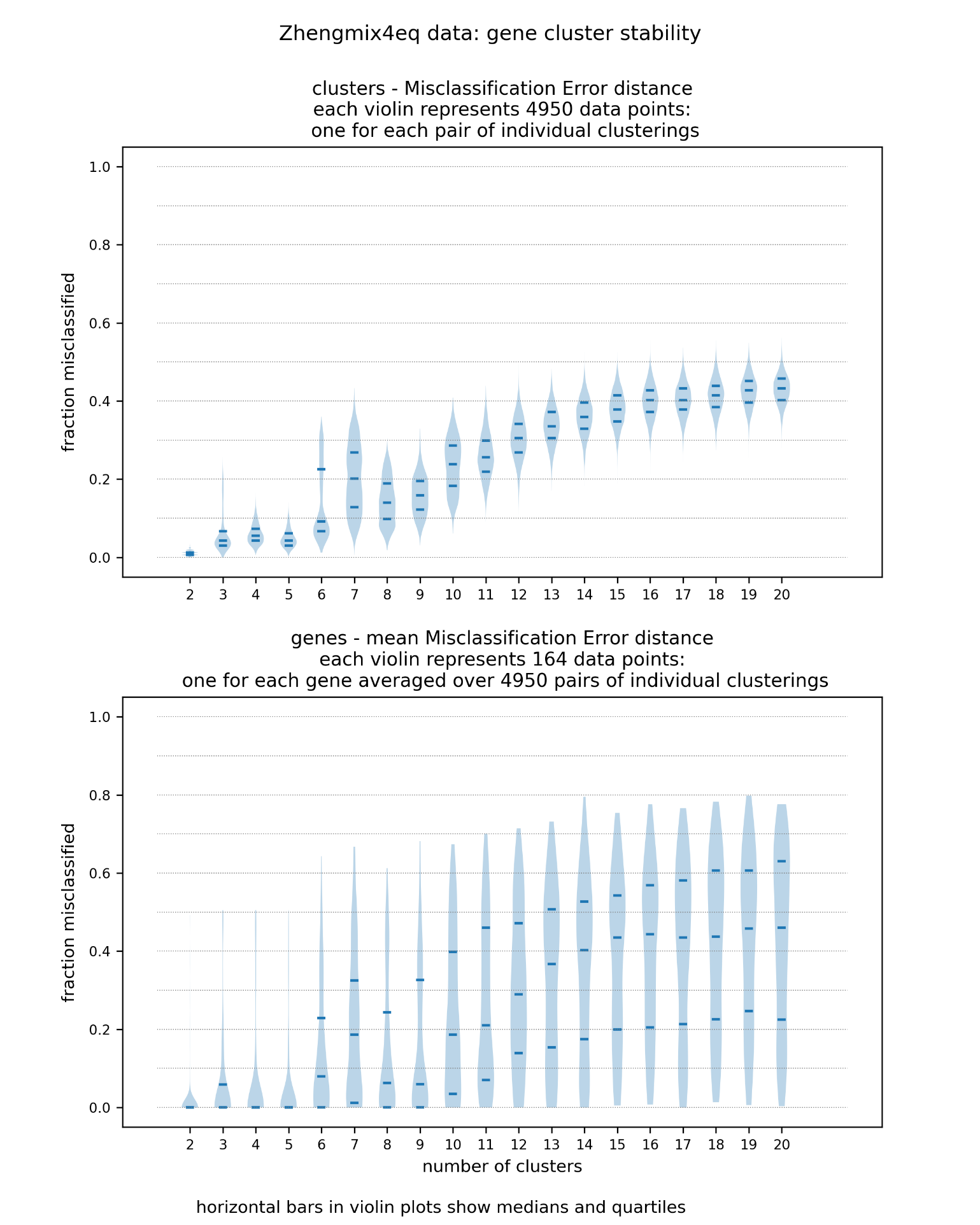


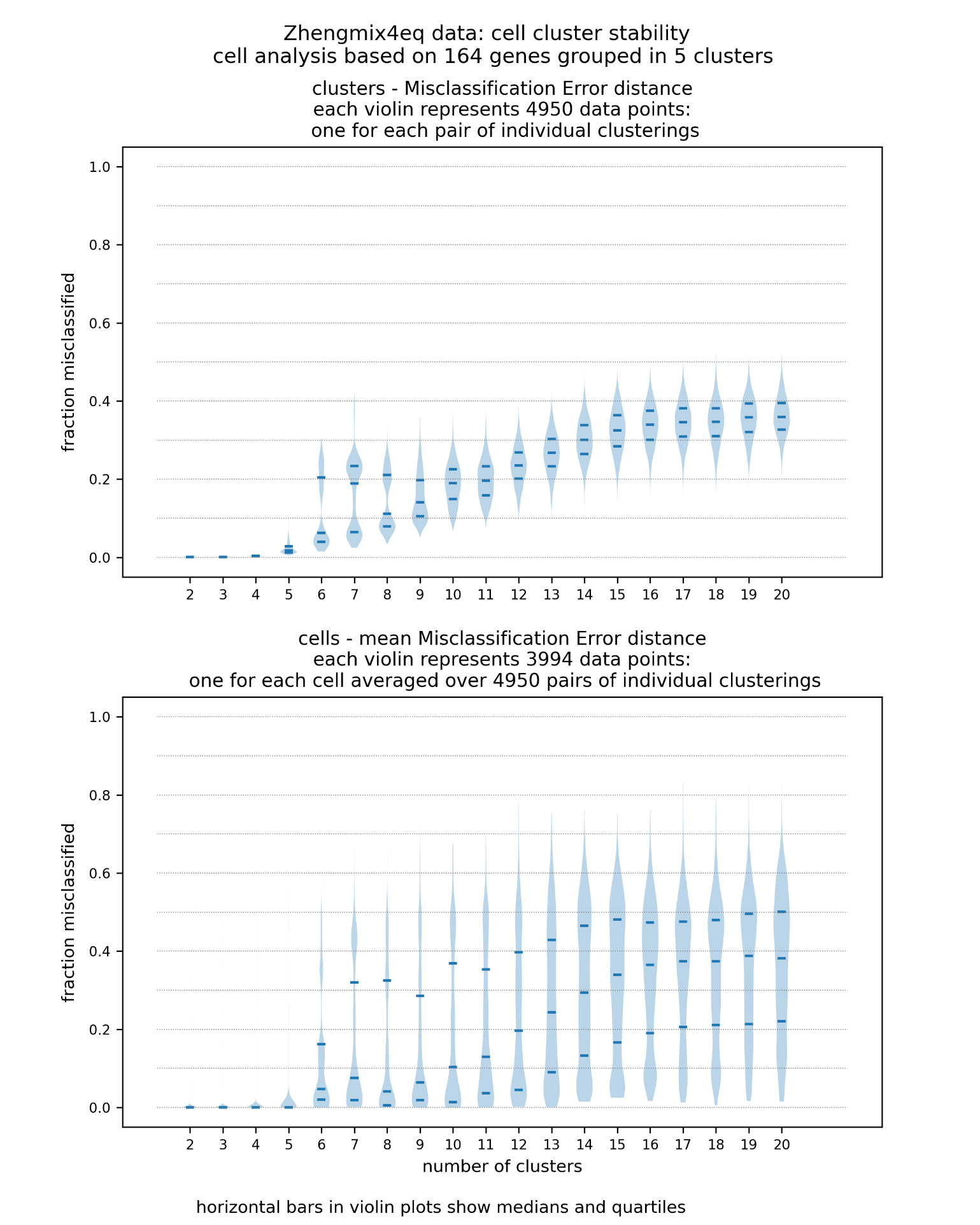
Figure 2 

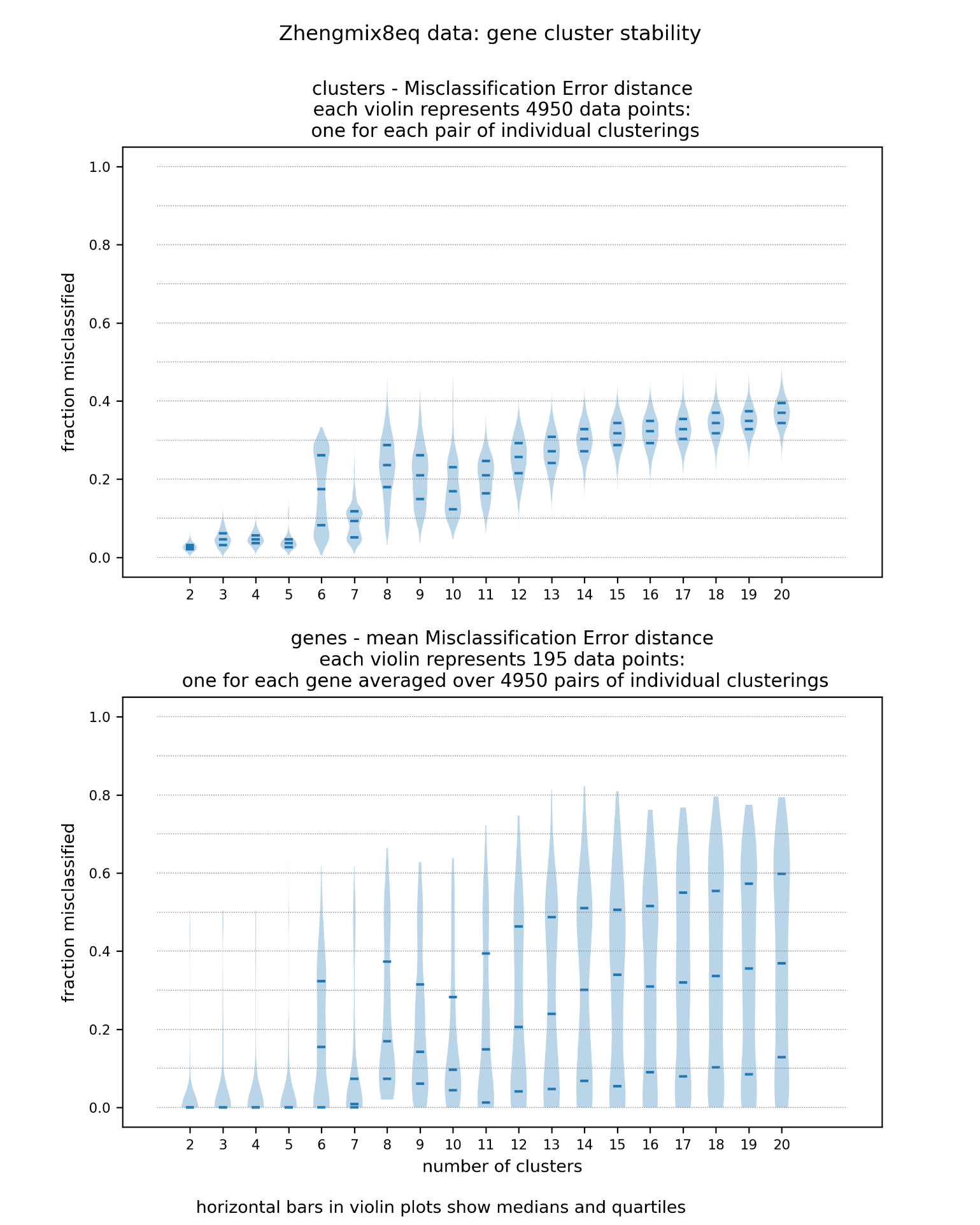
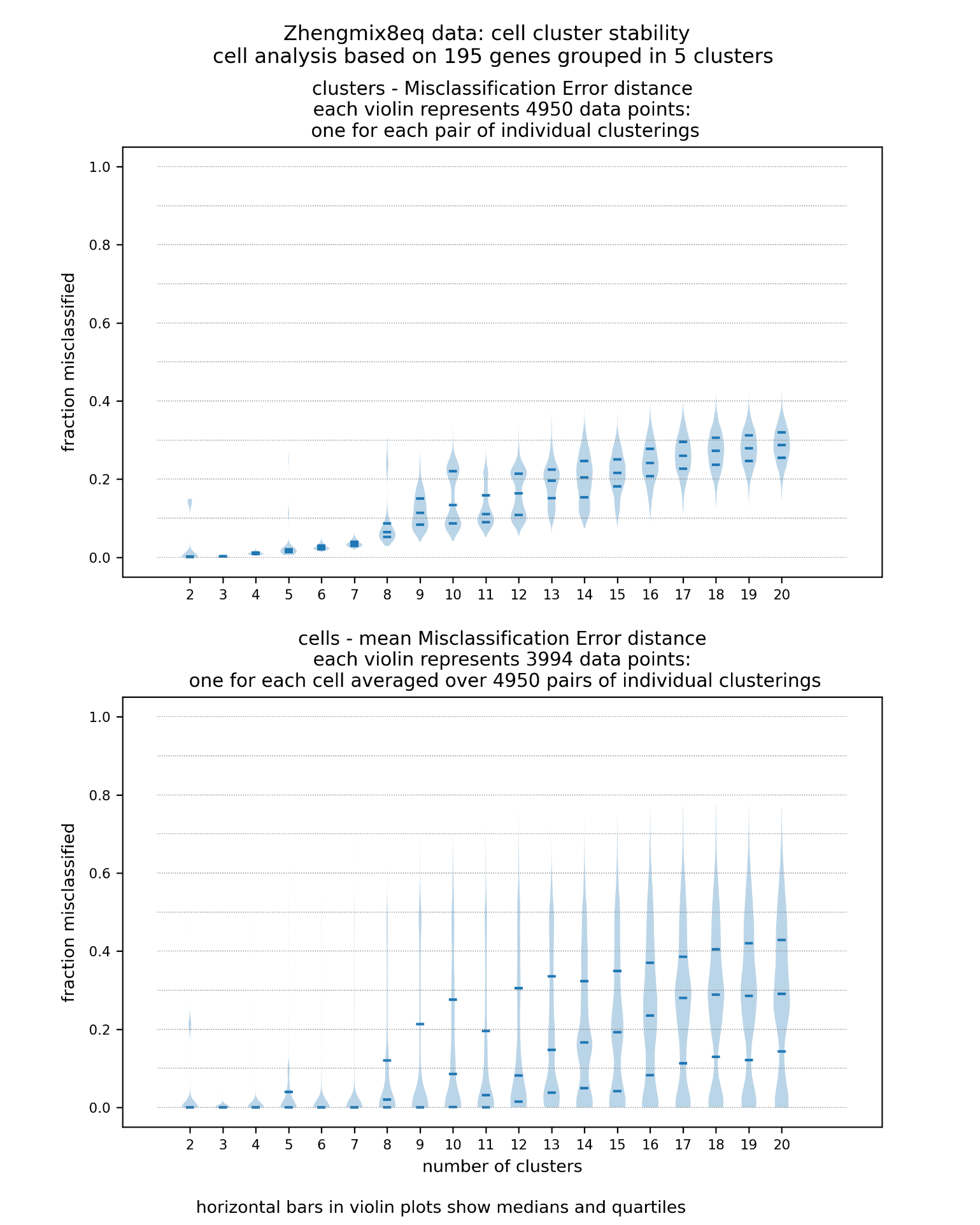
Figure 3

Figure 4



**References**

1. Townes F W, Hicks S C, Aryee M J et al.: Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biol* 20, 295 (2019). <https://doi.org/10.118s13059-019-1861-6>
2. Breiman L, Cutler A: Random forests—Classification manual. *URL*  http://www.math.usu.edu/~adele/forests/
3. Shi T, Horvath S: Unsupervised learning with random forest predictors. *Journal of Computational and Graphical Statistics*2006, **15:**118-138.
4. Hastie T, Tibshirani R, Friedman J: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition; section 14.2.4 “Unsupervised as Supervised Learning”
5. Ng A, Jordan M, Weiss Y: On spectral clustering: analysis and an algorithm. In T. Dietterich, S. Becker, and Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems* 14 (pp. 849 – 856). MIT Press, Cambridge (2002)
6. Meila M: The stability of a good clustering. <https://stat.washington.edu/sites/default/files/files/reports/2014/tr624.pdf>
7. He X, Cai D, Niyogi P: Laplacian score for feature selection. In *Advances in neural information processing systems*. 2006: 507-514.
8. Duo, A, Soneson, C: Package DuoClustering2018, Bioconductor
9. Ding C: Analysis of gene expression profiles: class discovery and leaf ordering. Proc. Conf. Research in *Comp.Mol.Bio* (RECOMB 2002), pp.127-136. April 2002, Washington, DC.
10. Tibshirani R, Walther G, Hastie T: Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society*: Series B (Statistical Methodology). 63 (2): 411–423. doi:10.1111/1467-9868.00293. ISSN 1467-9868.
11. Dudoit S, Fridlyand J: A prediction-based resampling method for estimating the number of clusters in a dataset. Genome Biol 3, research0036.1 (2002). <https://doi.org/10.1186/gb-2002-3-7-research0036>
12. Tibshirani R, Walther G: Cluster validation by prediction strength, *Journal of Computational and Graphical Statistic* 2005, 14:3, 511-528, DOI: 10.1198/106186005X59243