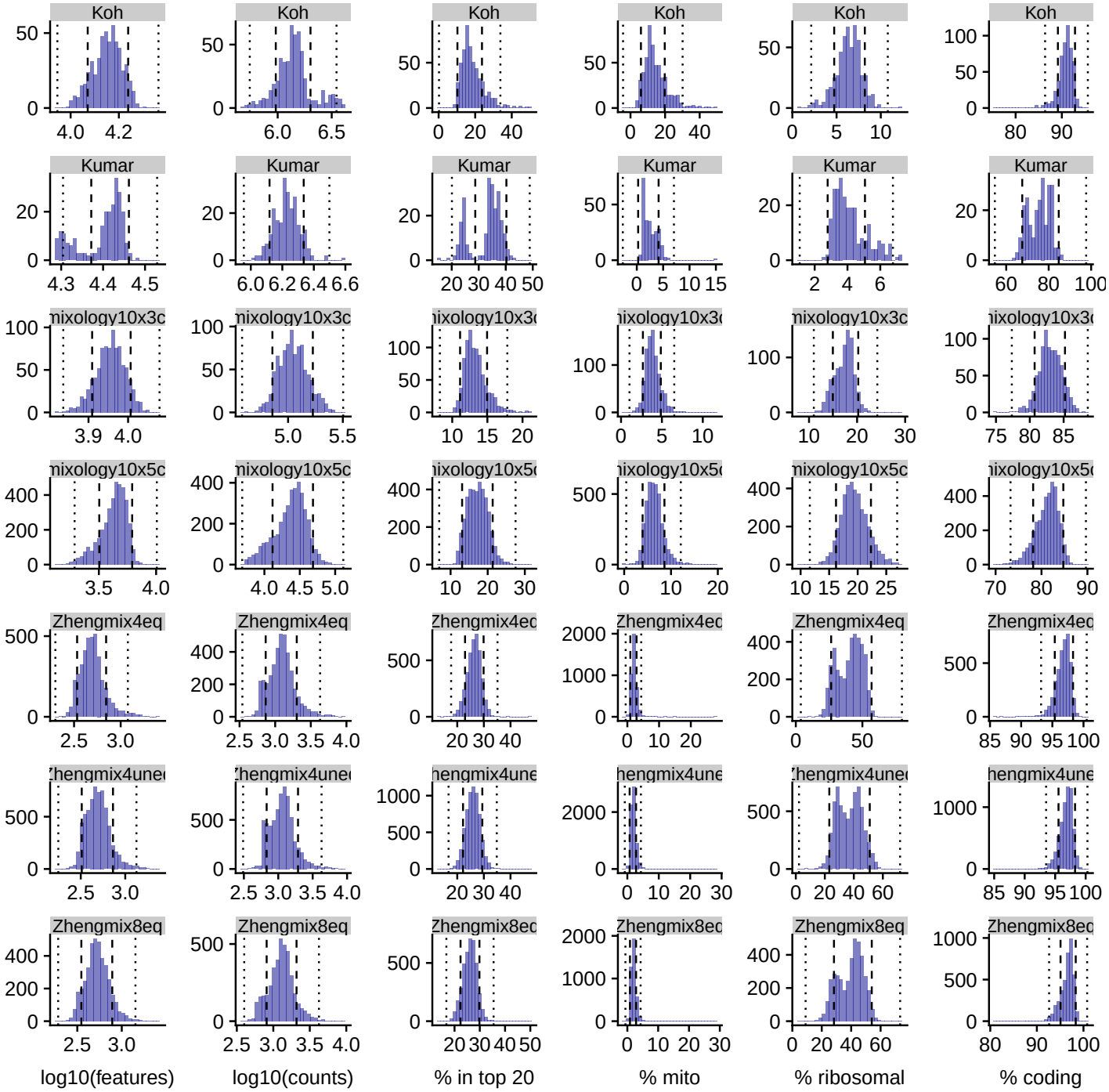
Supplementary Figures

Pierre-Luc Germain

18 Mai, 2019

Contents

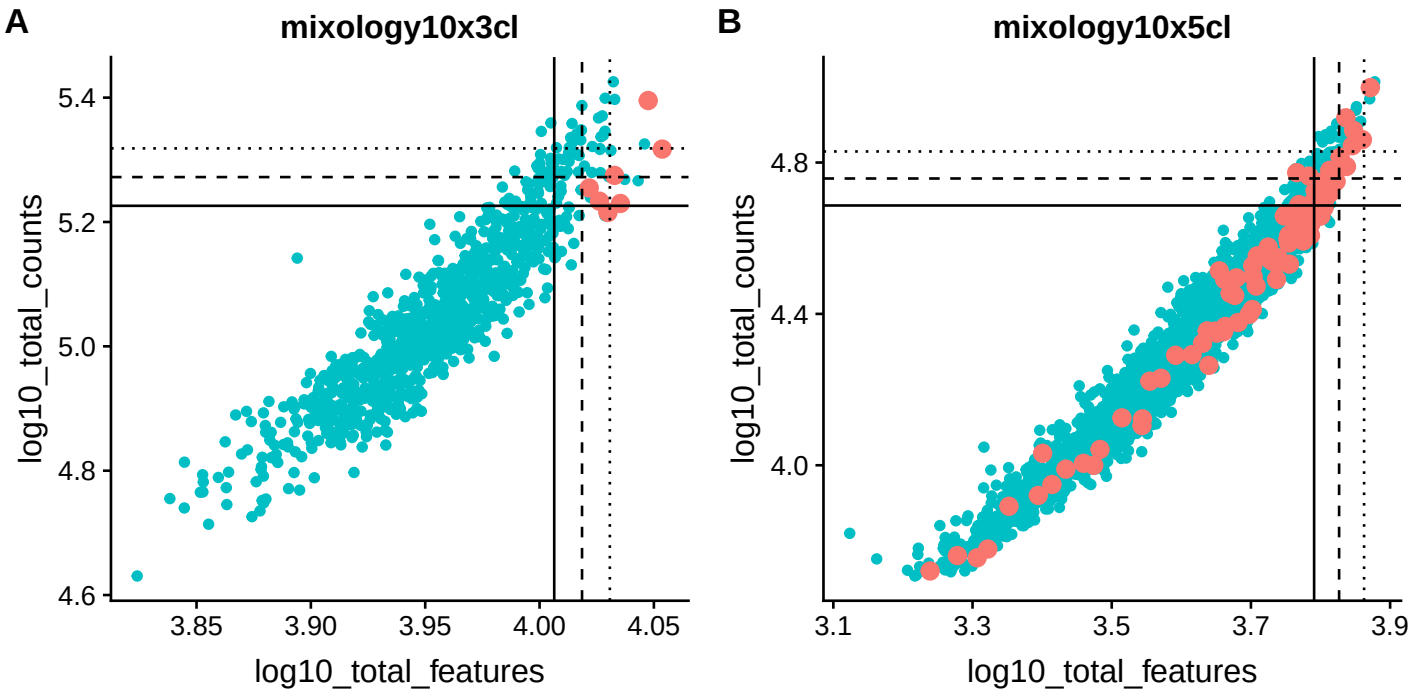
Supplementary Figure 1



Supplementary Figure 1

Distribution across cells of various control properties in the different datasets. The lines indicate respectively 2 and 5 median absolute deviations (MADs).

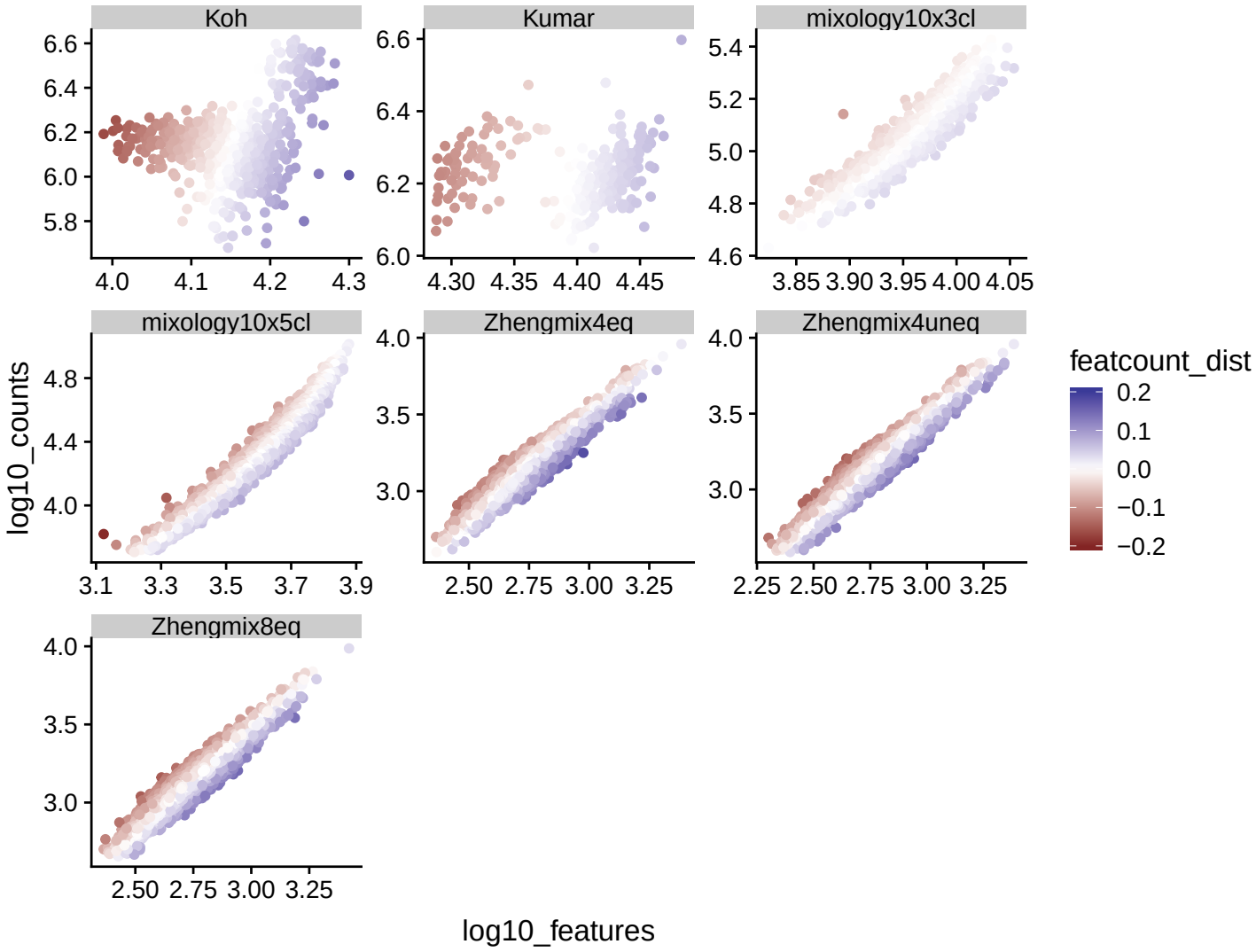
Supplementary Figure 2



Supplementary Figure 2

The total counts and total features per cell of doublets (red) versus other cells. We used the demuxlet annotation of doublets (based on SNPs) made available through CellBench. The lines indicate, respectively, 2, 2.5, and 3 median absolute deviations. While doublets tend to have a higher total count and especially number of detected features, these features alone are not always sufficient for their identification.

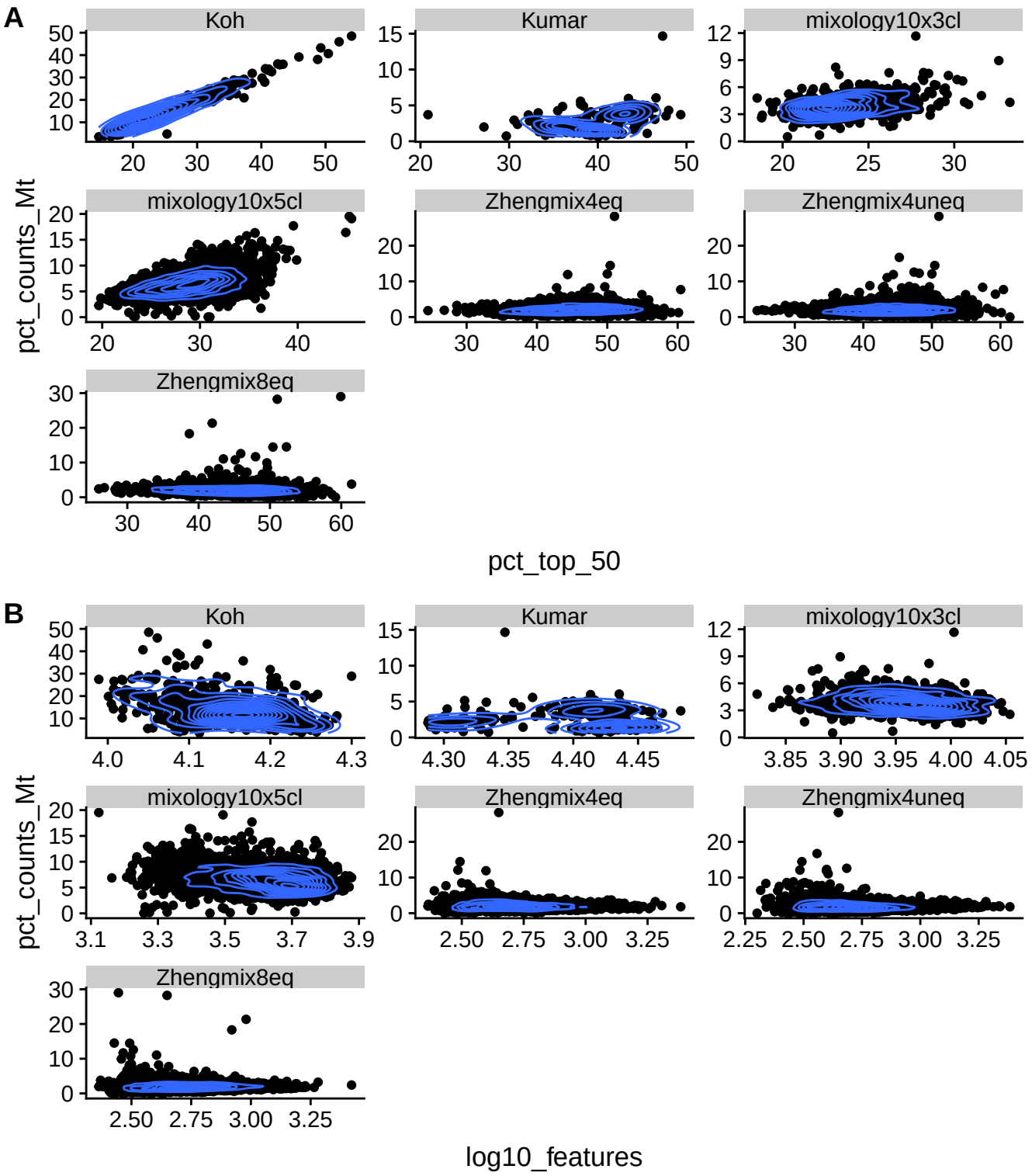
Supplementary Figure 3



Supplementary Figure 3

There is a tight relationship, in 10x datasets (i.e. not the `Koh` and `Kumar` datasets), between the total counts of a cell and its number of detected features. We therefore include, among control variables, deviation from this ratio.

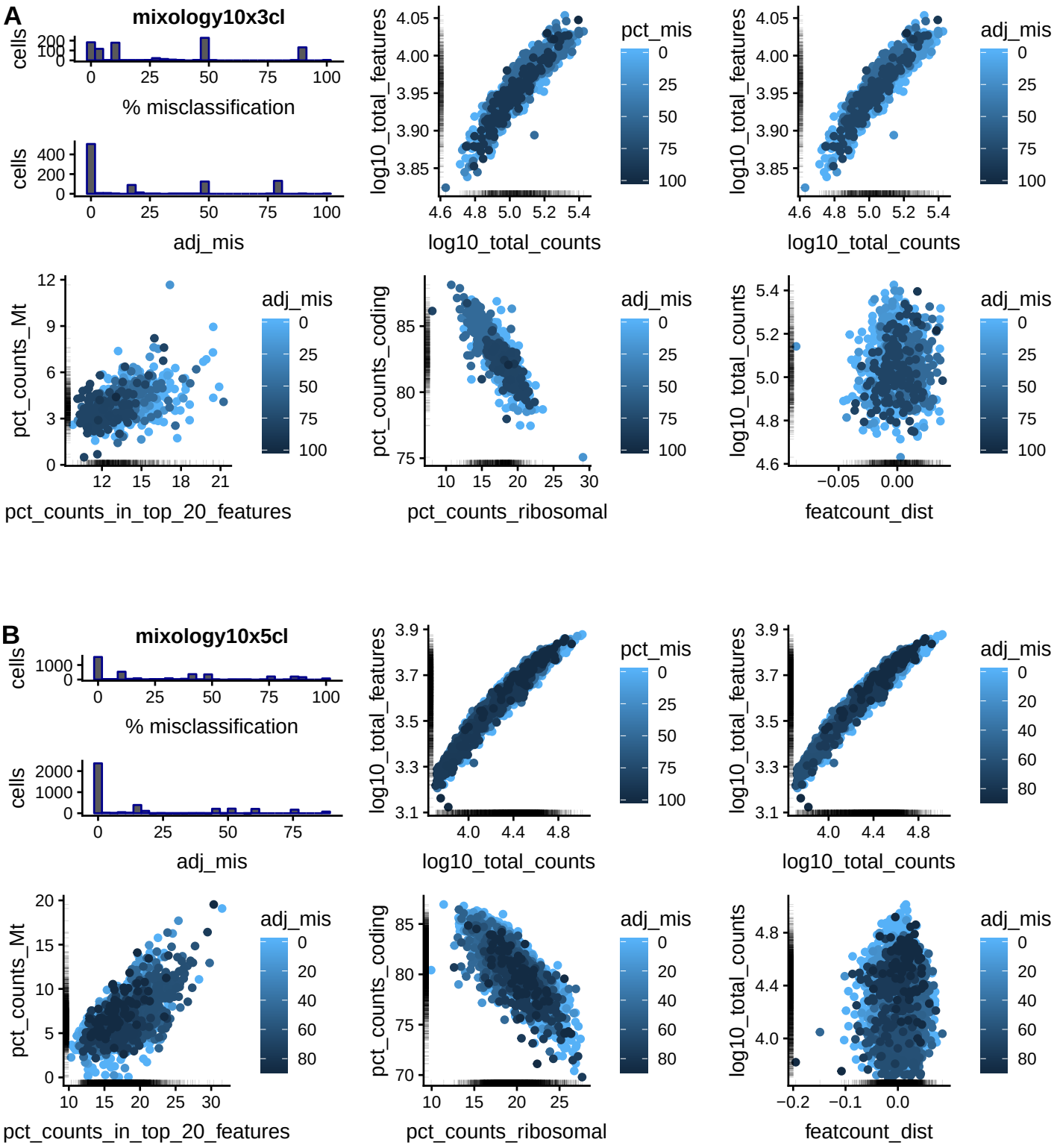
Supplementary Figure 4



Supplementary Figure 4

There is a tight relationship, in 10x datasets (i.e. not the `Koh` and `Kumar` datasets), between the total counts of a cell and its number of detected features. We therefore include, among control variables, deviation from this ratio.

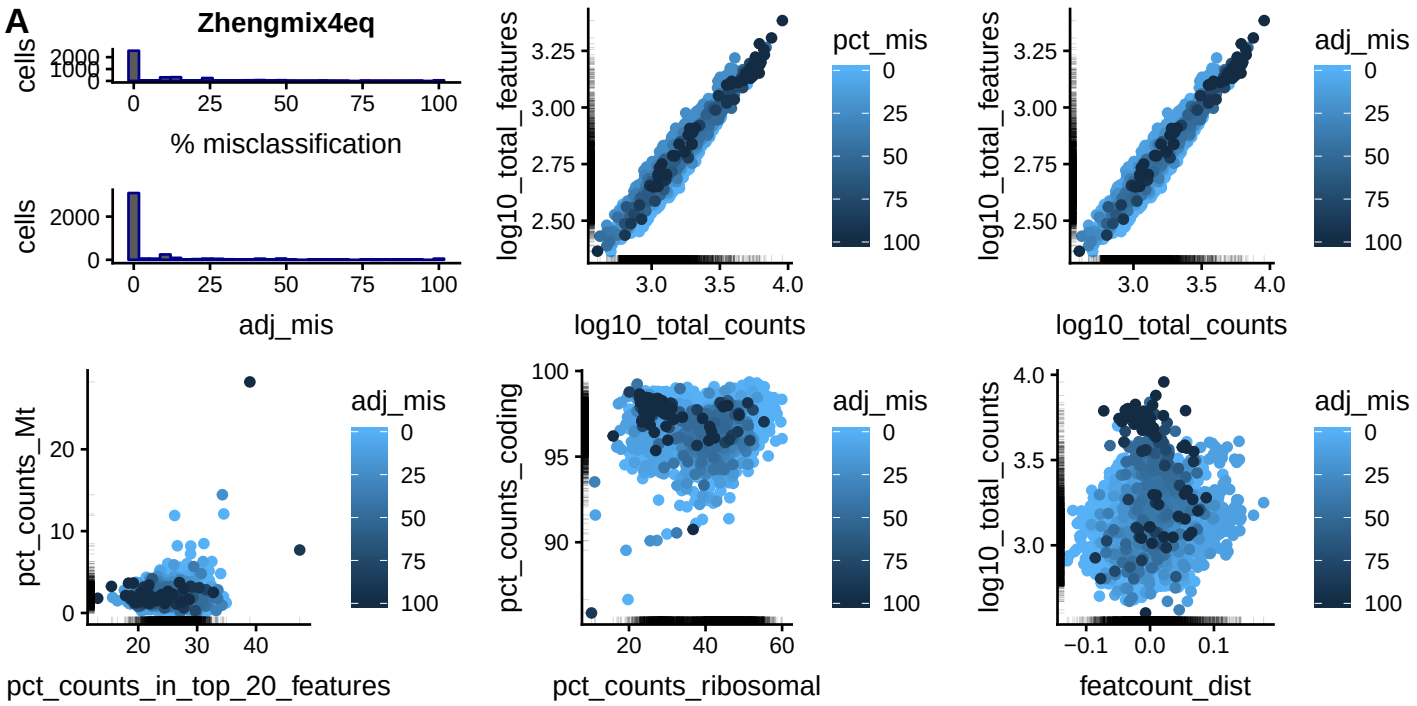
Supplementary Figure 5

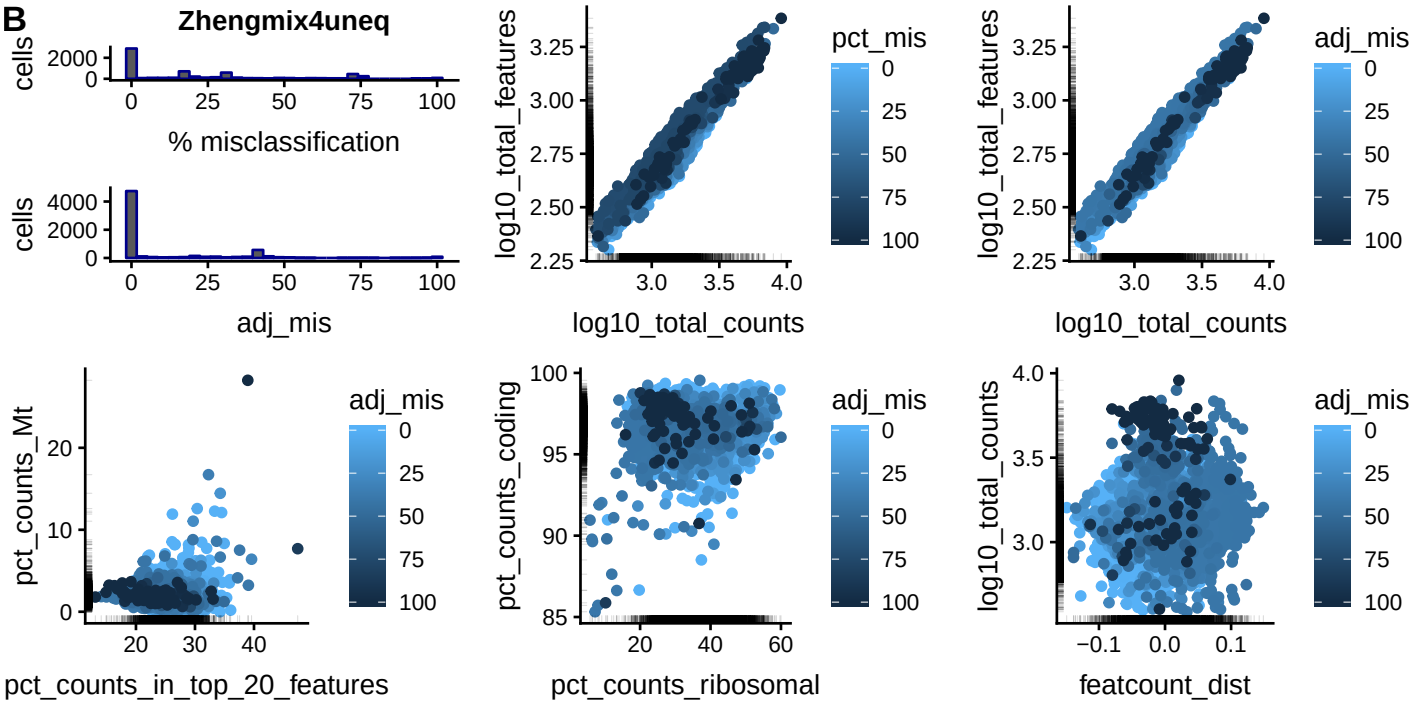


Supplementary Figure 5

Relationship between various cellular properties and the frequency of cluster mis-assignment for the mixology10x3cl (A) and mixology10x5cl (B) datasets. The percentage of misclassification refers to the frequency with which a given cell is assigned the wrong cluster (using the Hungarian algorithm for cluster matching) across several hundred clustering runs with varying parameters. Since some subpopulations tend to be more misclassified than others, the adjusted rate of misclassification (`adj_mis`) is substracted for the subpopulation median misclassification rate.

Supplementary Figure 6

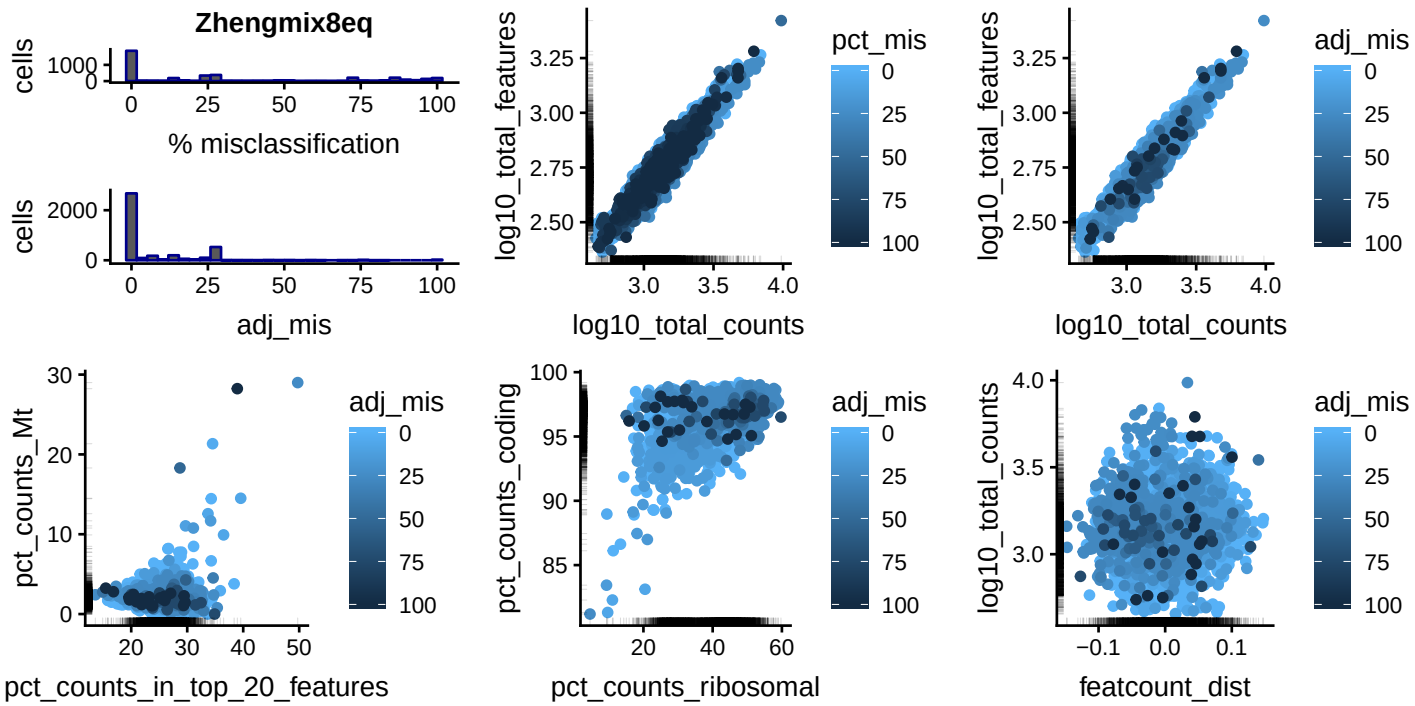




Supplementary Figure 6

Relationship between various cellular properties and the frequency of cluster mis-assignment for the Zheng equal (A) or unequal (B) mixtures of four cell types. See Supplementary Figure 5 for more information. The only clear pattern is that cells with a high number of reads or features tend to have a higher misclassification rate.

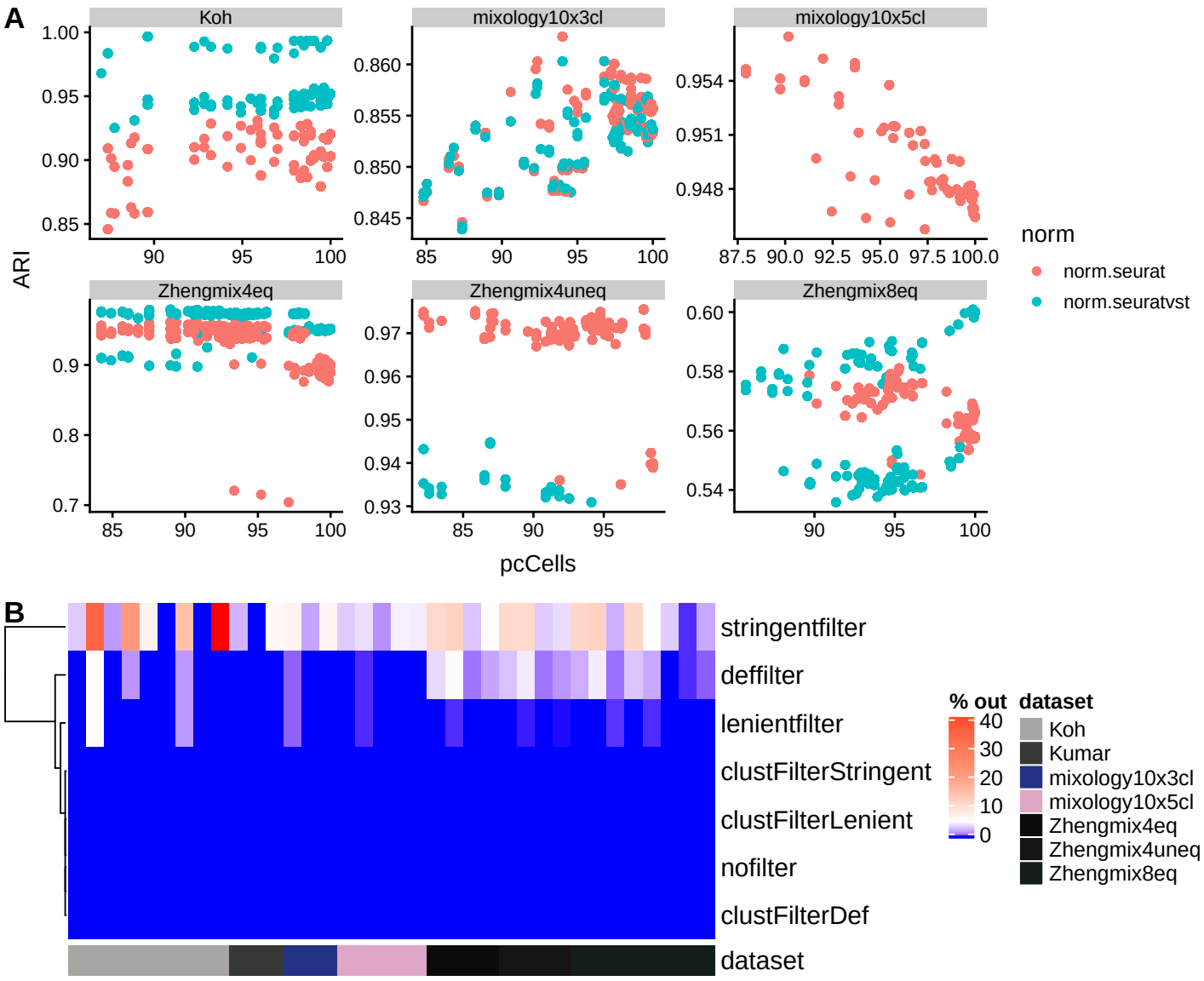
Supplementary Figure 7



Supplementary Figure 7

Relationship between various cellular properties and the frequency of cluster mis-assignment for the Zheng mixture of 8 cell types. See Supplementary Figure 5 for more information.

Supplementary Figure 8



Supplementary Figure 8

A: Adjusted rand index (ARI) of various clustering pipelines using various combinations of filtering criteria. Only clustering analyses with a number of clusters +/-1 from the real number of clusters were considered. In general, filtering more does not result in more accurate clustering. **B:** Percentage of cells filtered out per subpopulation, using various combination of filters (see methods).

Supplementary Figure 9

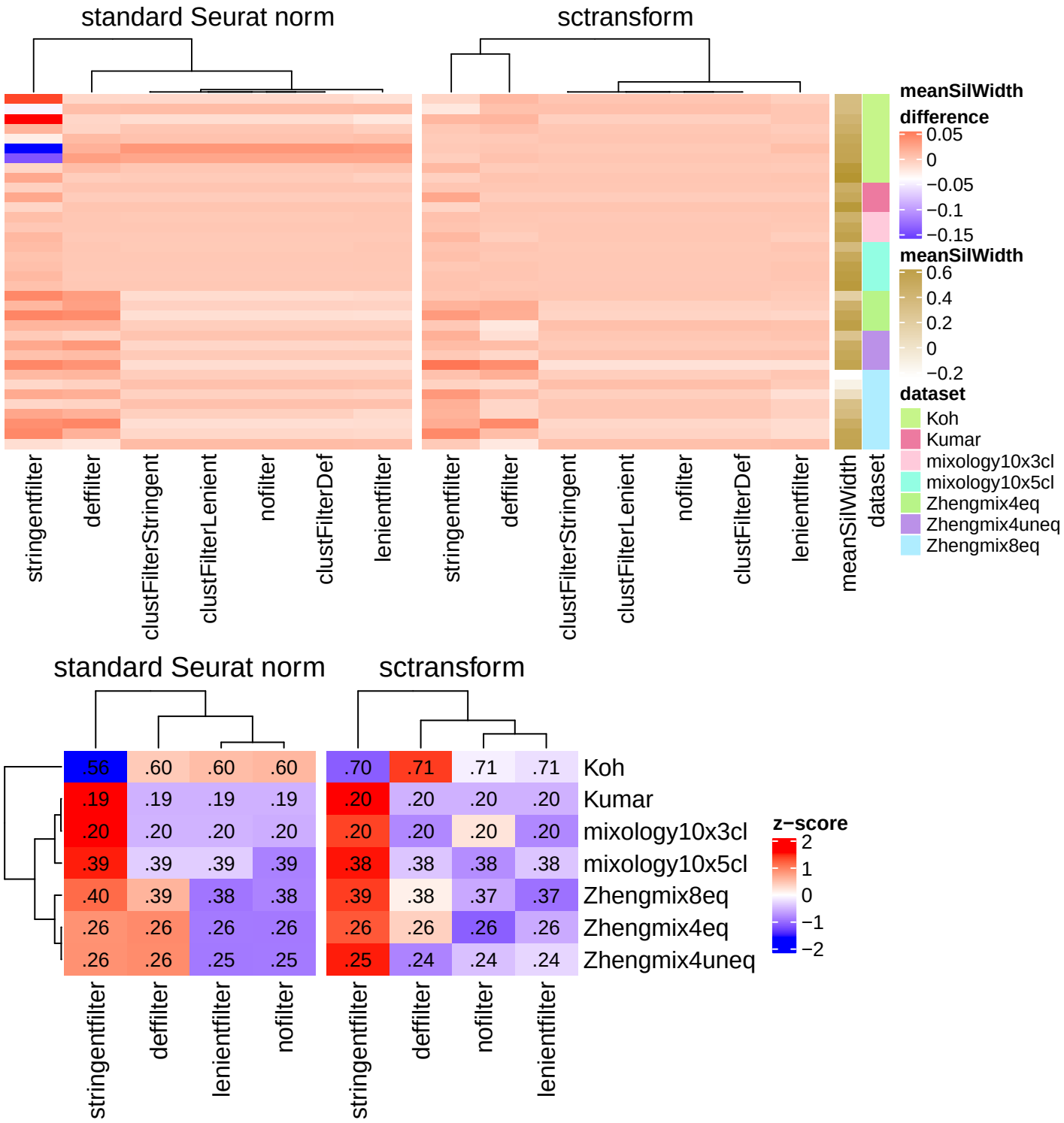
```
## lm(ARI~nbClusters*dataset+norm+factor(filt.times)+factor(filt.MADs)+counts.lower+counts.higher+mt.higher+f
```

	Estimate	Std. Error	t value	p value
normnorm.seuratvst	0.0069338	0.0003104	22.3415248	0.0000000
factor(filt.times)2	0.0086024	0.0003617	23.7847962	0.0000000
factor(filt.times)3	0.0168470	0.0003811	44.2112922	0.0000000
factor(filt.MADs)2.5	-0.0040702	0.0004199	-9.6944075	0.0000000
factor(filt.MADs)3	-0.0054701	0.0004202	-13.0177667	0.0000000
factor(filt.MADs)5	-0.0025099	0.0004205	-5.9684908	0.0000000
counts.lowerTRUE	-0.0001193	0.0003534	-0.3375833	0.7356780
counts.higherTRUE	-0.0073031	0.0003552	-20.5607582	0.0000000
mt.higherTRUE	0.0006318	0.0007765	0.8136392	0.4158539
feat.lowerTRUE	0.0002074	0.0003494	0.5934348	0.5528917
feat.higherTRUE	-0.0058113	0.0003506	-16.5729653	0.0000000
top50.lowerTRUE	-0.0000717	0.0003045	-0.2354817	0.8138354
top50.higherTRUE	-0.0000717	0.0003045	-0.2354817	0.8138354
ratiodist.lowerTRUE	-0.0009712	0.0003045	-3.1895182	0.0014256
ratiodist.higherTRUE	0.0012337	0.0003045	4.0515679	0.0000509

Supplementary Figure 9

Linear regression analysis of ARI on various filtering criteria, correcting for the effect of the number of clusters. Increasing the number of distributions on which a cell must deviate in order to be excluded (`filt.times`) tended to be positively associated with ARI, while increasing the number of median absolute deviations required to be an outlier (`filt.MADs`) was associated with a decrease in ARI.

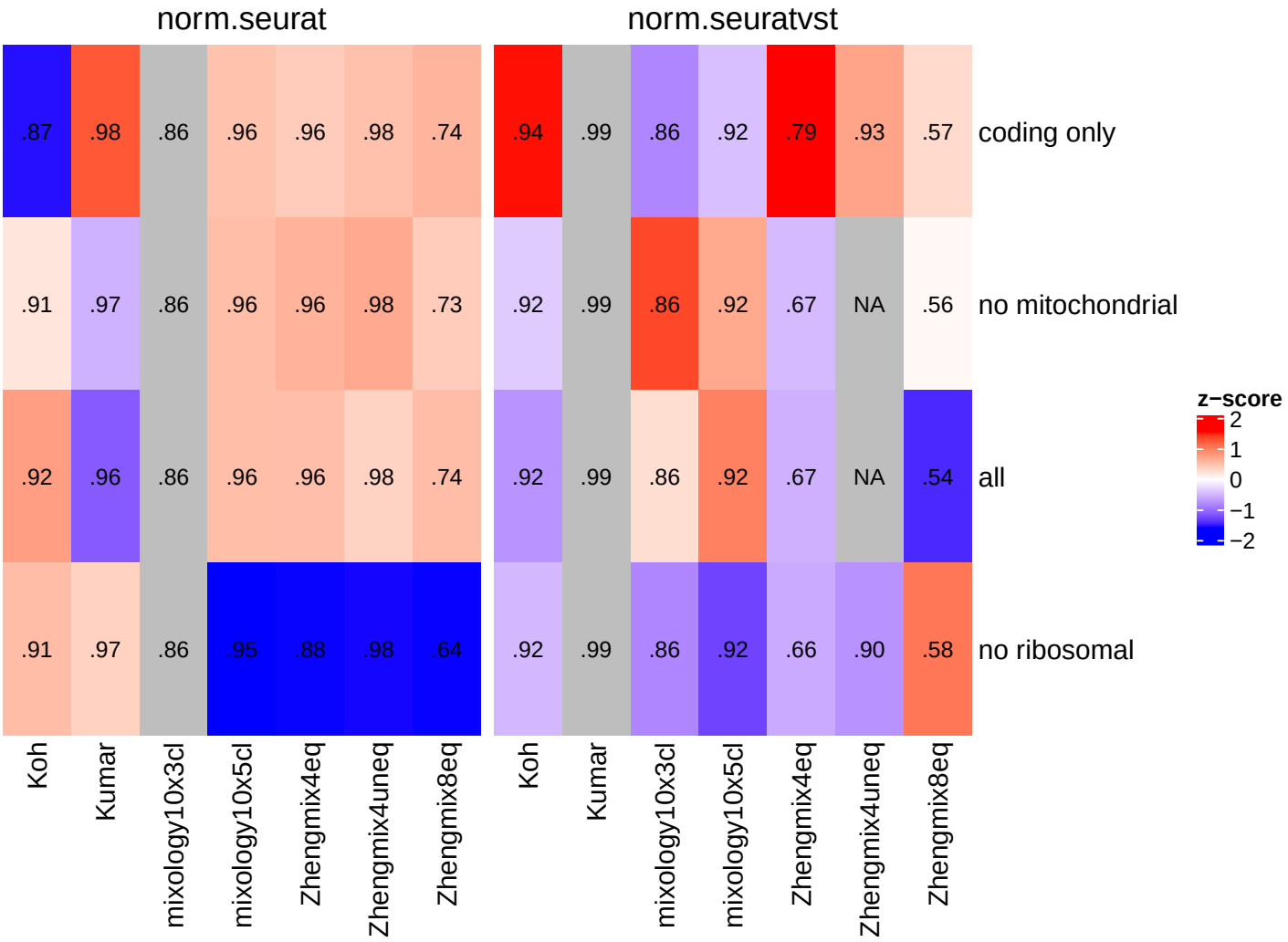
Supplementary Figure 10



Supplementary Figure 10

Impact of the set of filters on the average silhouette width of the subpopulations (top) and on the proportion of variance in the first components explained by clusters (bottom).

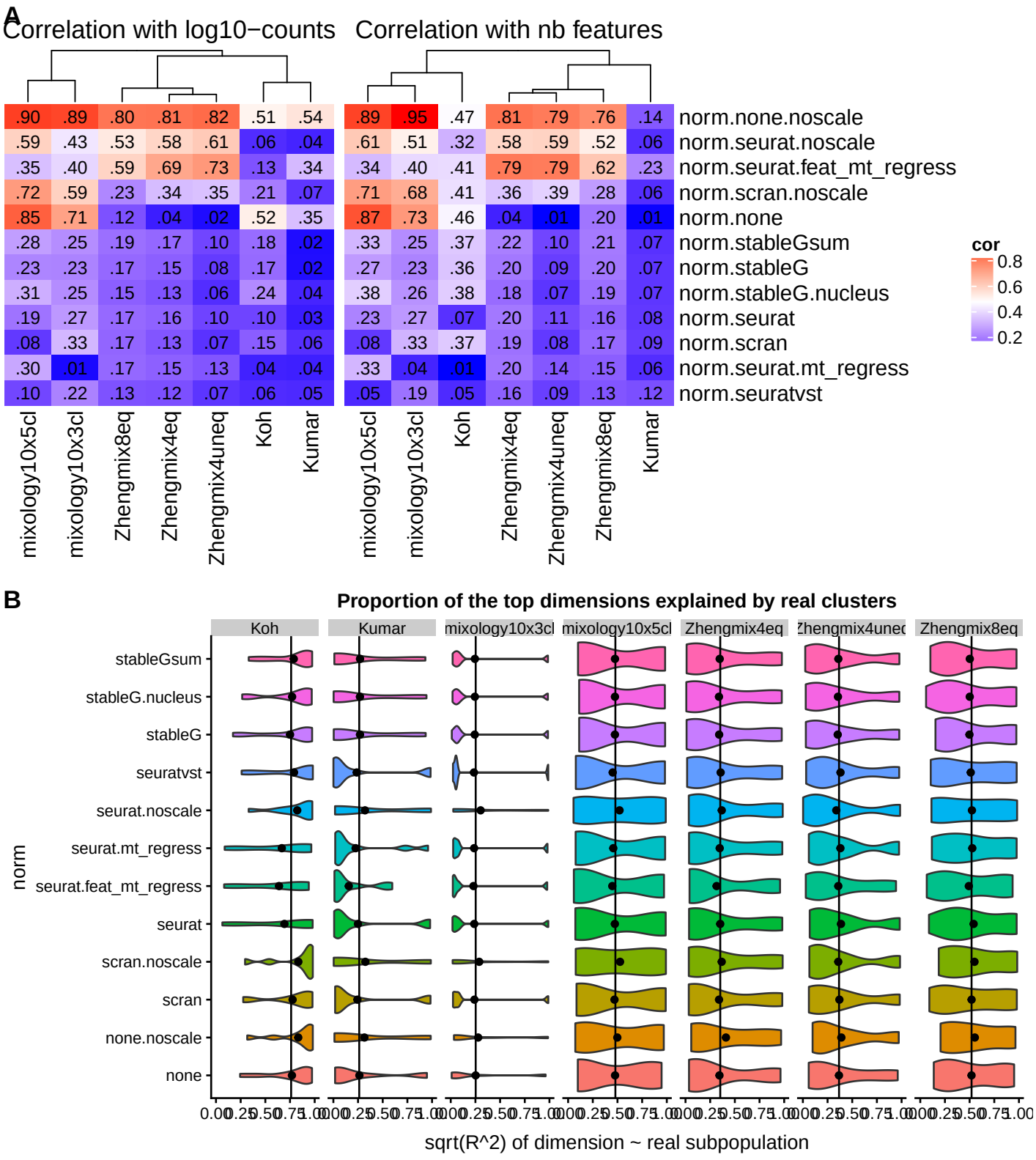
Supplementary Figure 11



Supplementary Figure 11

Impact of restricting the type of features used on the ARI of the clustering.

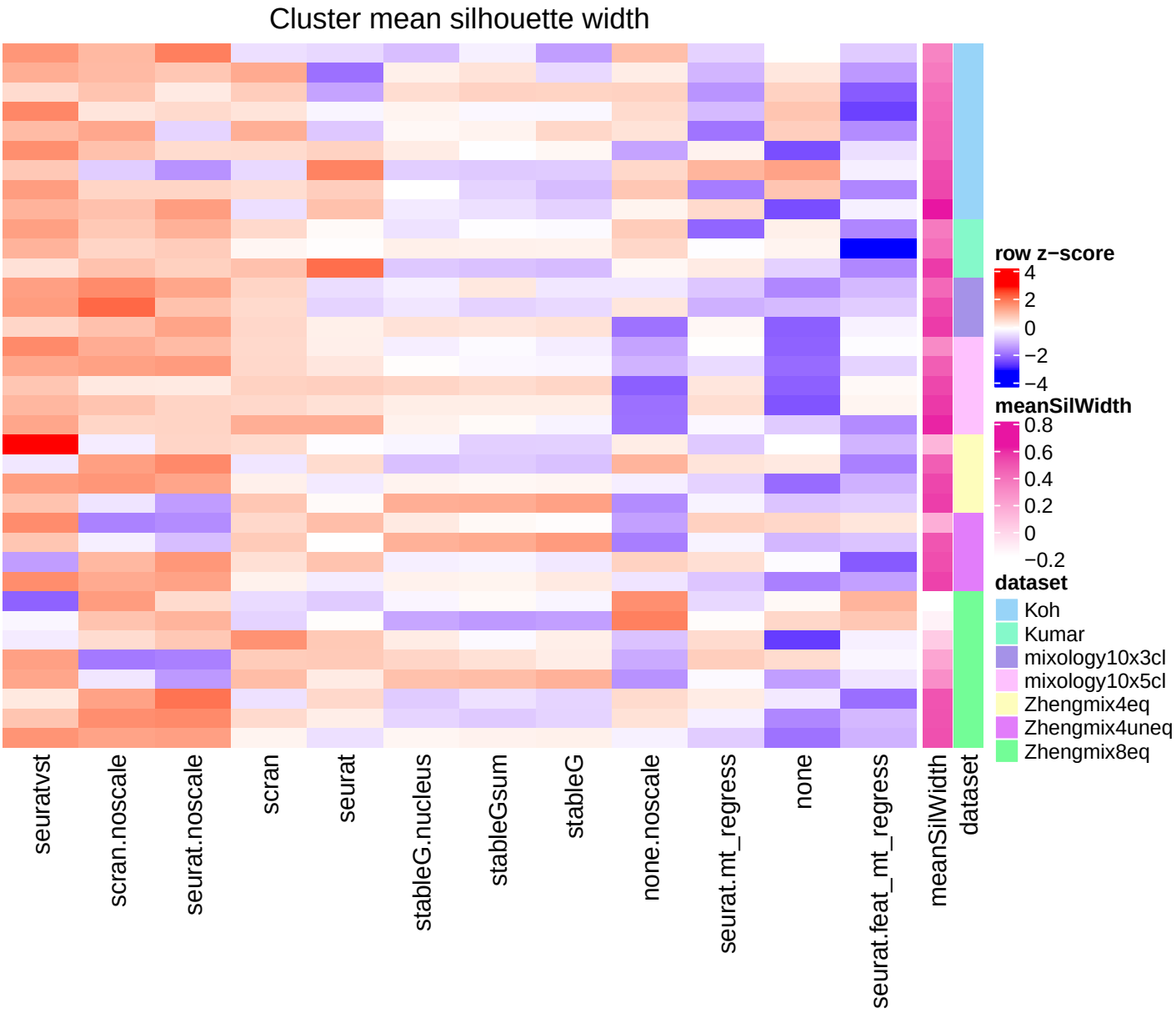
Supplementary Figure 12



Supplementary Figure 12

A: Correlation between library size (left) or the number of detected features (right) and the residuals of a linear regression of the first principal component on the subpopulation. **B:** Proportion of the variance in the first 10 PCA components that is explained by real subpopulations. The square-root of R-squared values is plotted for ease of visualization.

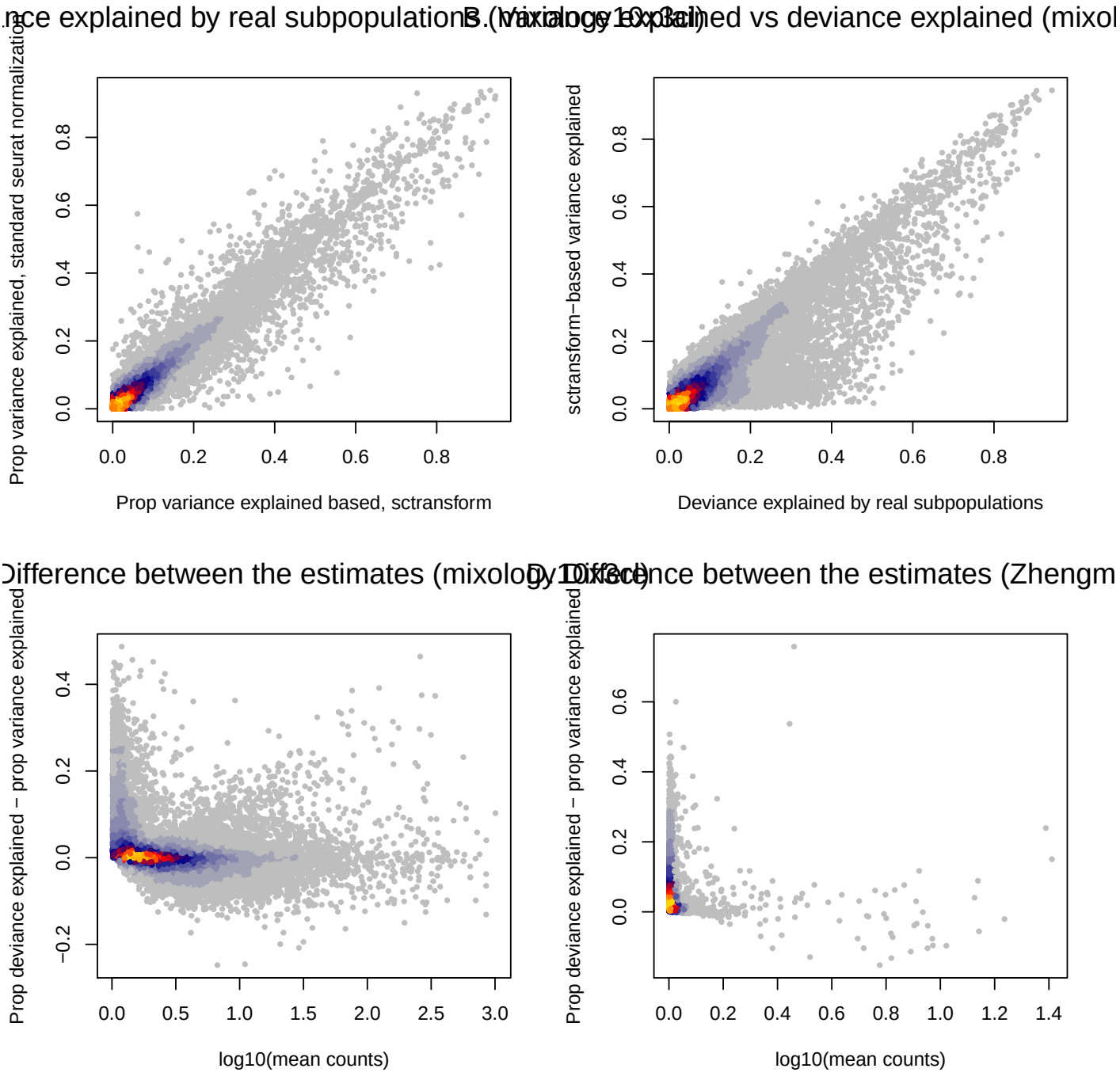
Supplementary Figure 13



Supplementary Figure 13

Mean silhouette width of true subpopulations across the first 10 components with different normalization and scaling approaches.

Supplementary Figure 14



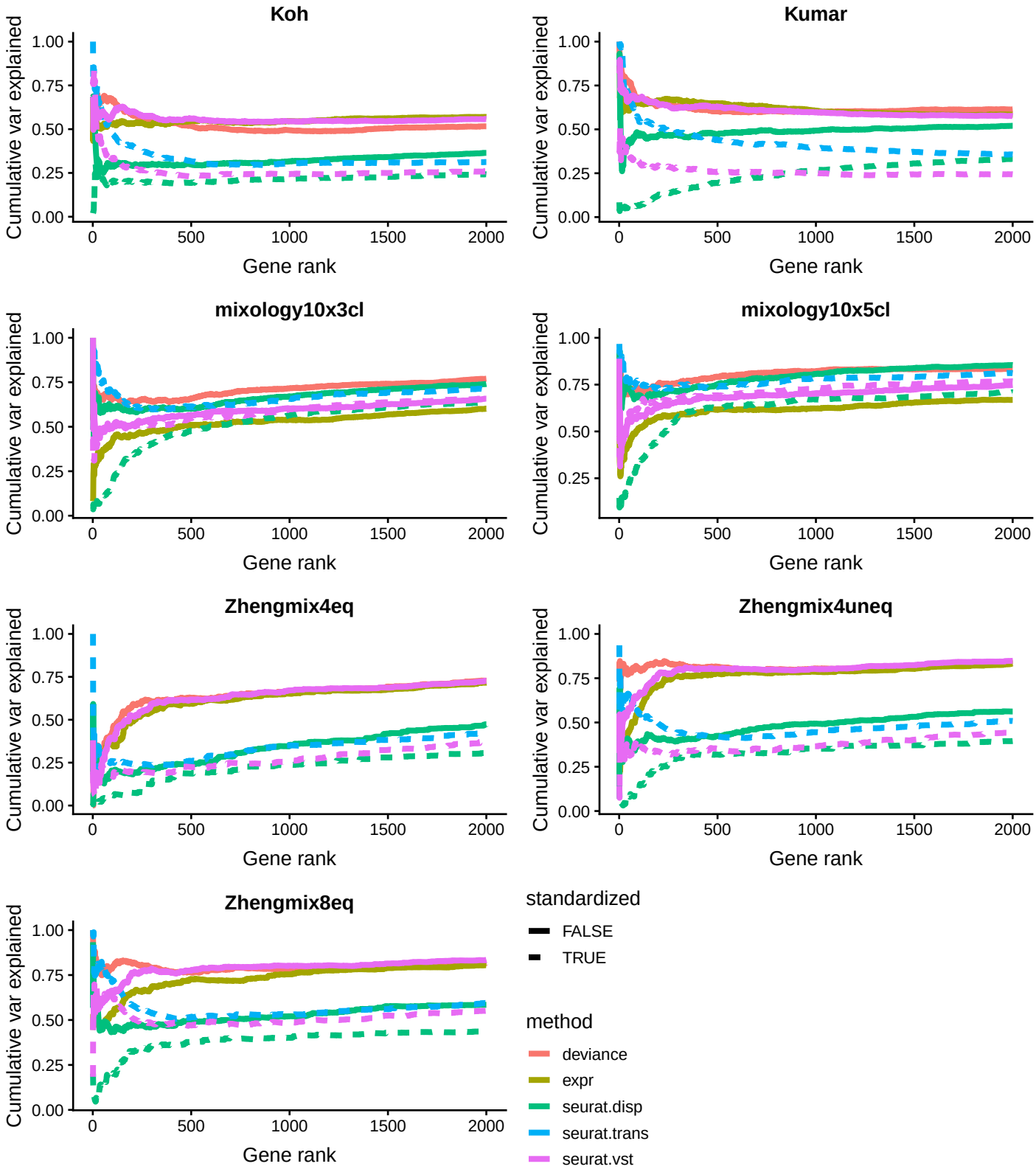
Supplementary Figure 14

A: Comparison of the gene-wise proportion of variance explained by real subpopulations based on Seurat's standard log normalization and on `sctransform` variance-stabilizing transformation. Across 10x datasets, there is a good agreement between the two, the correlation ranging between 0.92 and 0.97. **B:** There is also a good agreement between *variance* and *deviance* explained, with some genes having a higher deviance explained. **C-D:** Relationship between mean expression and the difference between the proportion of deviance explained and the proportion of variance explained in two datasets. Genes that have a higher proportion of the deviance explained than of the variance explained are generally the lowly-expressed ones.

Supplementary Figure 15

```
## Warning: Removed 328867 rows containing missing values (geom_path).
```

```
## Warning: Removed 328867 rows containing missing values (geom_path).
## Warning: Removed 302113 rows containing missing values (geom_path).
## Warning: Removed 101276 rows containing missing values (geom_path).
## Warning: Removed 68502 rows containing missing values (geom_path).
## Warning: Removed 94976 rows containing missing values (geom_path).
## Warning: Removed 101101 rows containing missing values (geom_path).
## Warning: Removed 96012 rows containing missing values (geom_path).
```



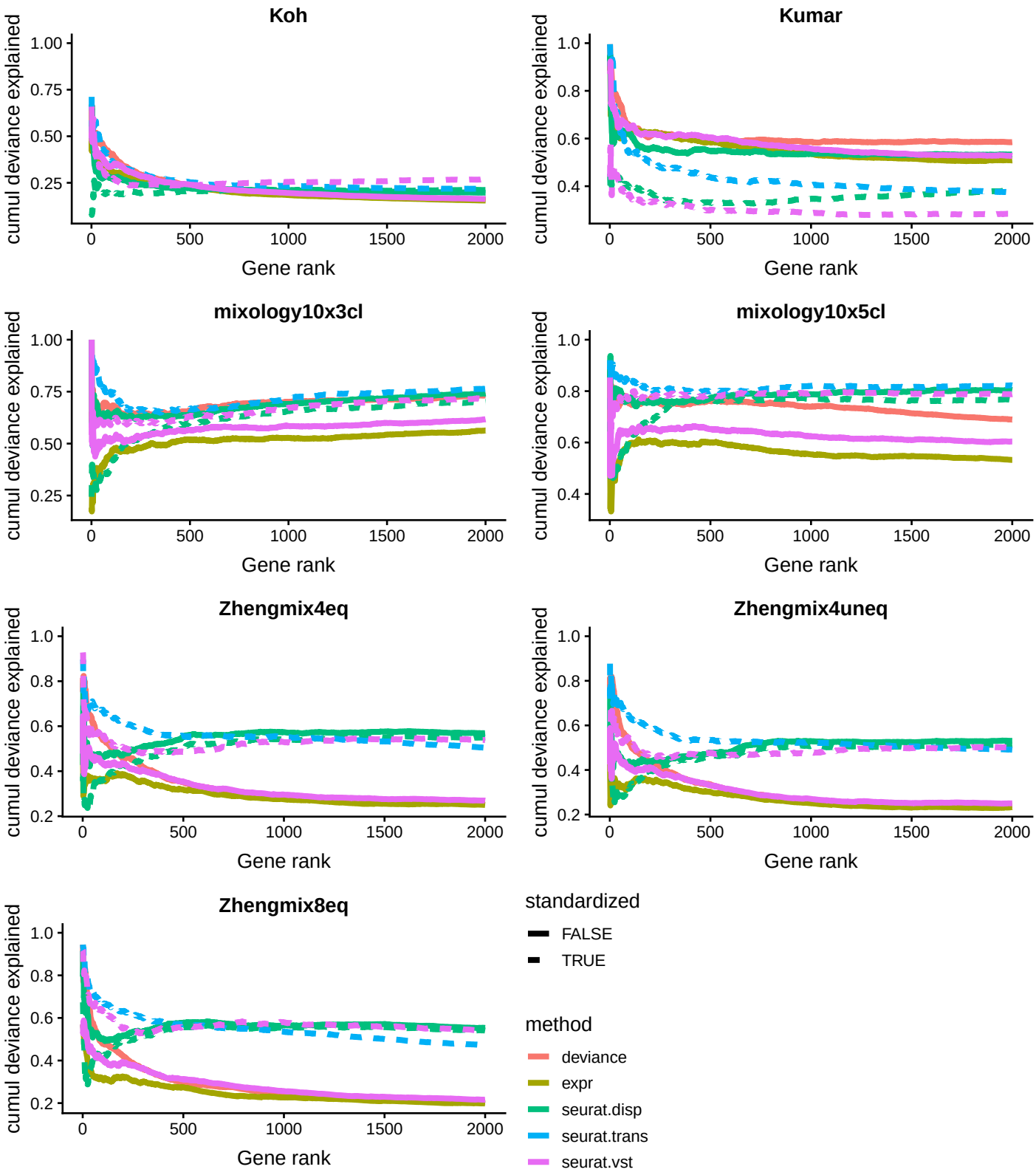
Supplementary Figure 15

Proportion of the cumulative *variance* explained by real subpulations that is retrieved through the selection. For each gene, we compute the proportion of the variance explained by real subpopulations. For each rank X, we sum this proportion for the X genes selected by a given method, and divide it by the sum when selecting the X genes with the highest variance explained. An ideal selection would therefore be a horizontal line at 1.

Supplementary Figure 16

```
## Warning: Removed 328867 rows containing missing values (geom_path).
```

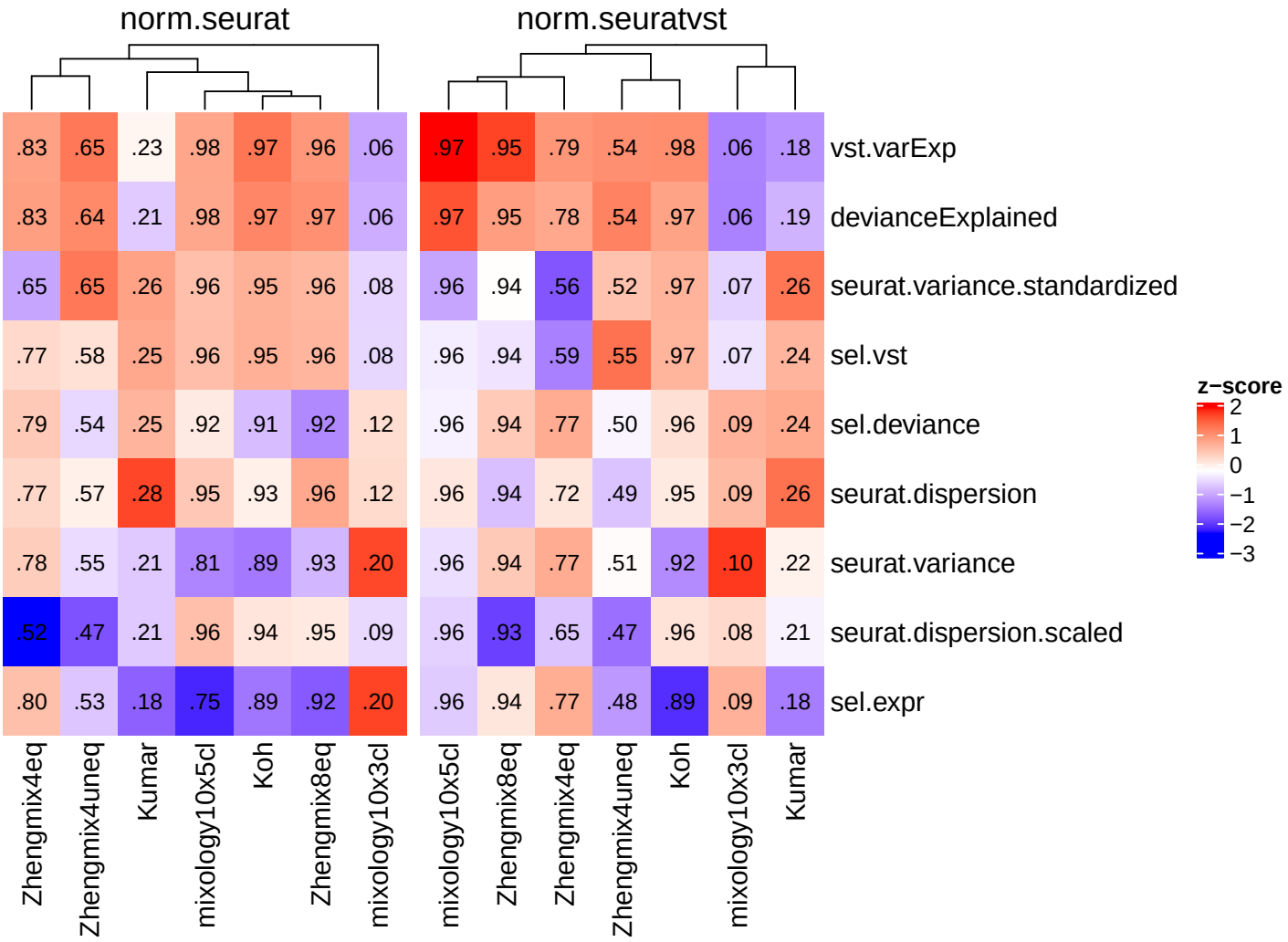
```
## Warning: Removed 328867 rows containing missing values (geom_path).
## Warning: Removed 302113 rows containing missing values (geom_path).
## Warning: Removed 101276 rows containing missing values (geom_path).
## Warning: Removed 68502 rows containing missing values (geom_path).
## Warning: Removed 94976 rows containing missing values (geom_path).
## Warning: Removed 101101 rows containing missing values (geom_path).
## Warning: Removed 96012 rows containing missing values (geom_path).
```



Supplementary Figure 16

Proportion of the cumulative *deviance* explained by real subpulations that is retrieved through the selection. For each gene, we compute the proportion of the variance explained by real subpopulations. As for Supplementary Figure 15, except using deviance explained.

Supplementary Figure 17



Supplementary Figure 17

Proportion of the variance in the first 5 principal components that is explained by real subpopulations according to different normalization and selection procedures, selecting 1000 genes in each case.