pipeComp, a general framework for the evaluation of computational pipelines, reveals performant single-cell RNA-seq preprocessing tools

Supplementary Figures

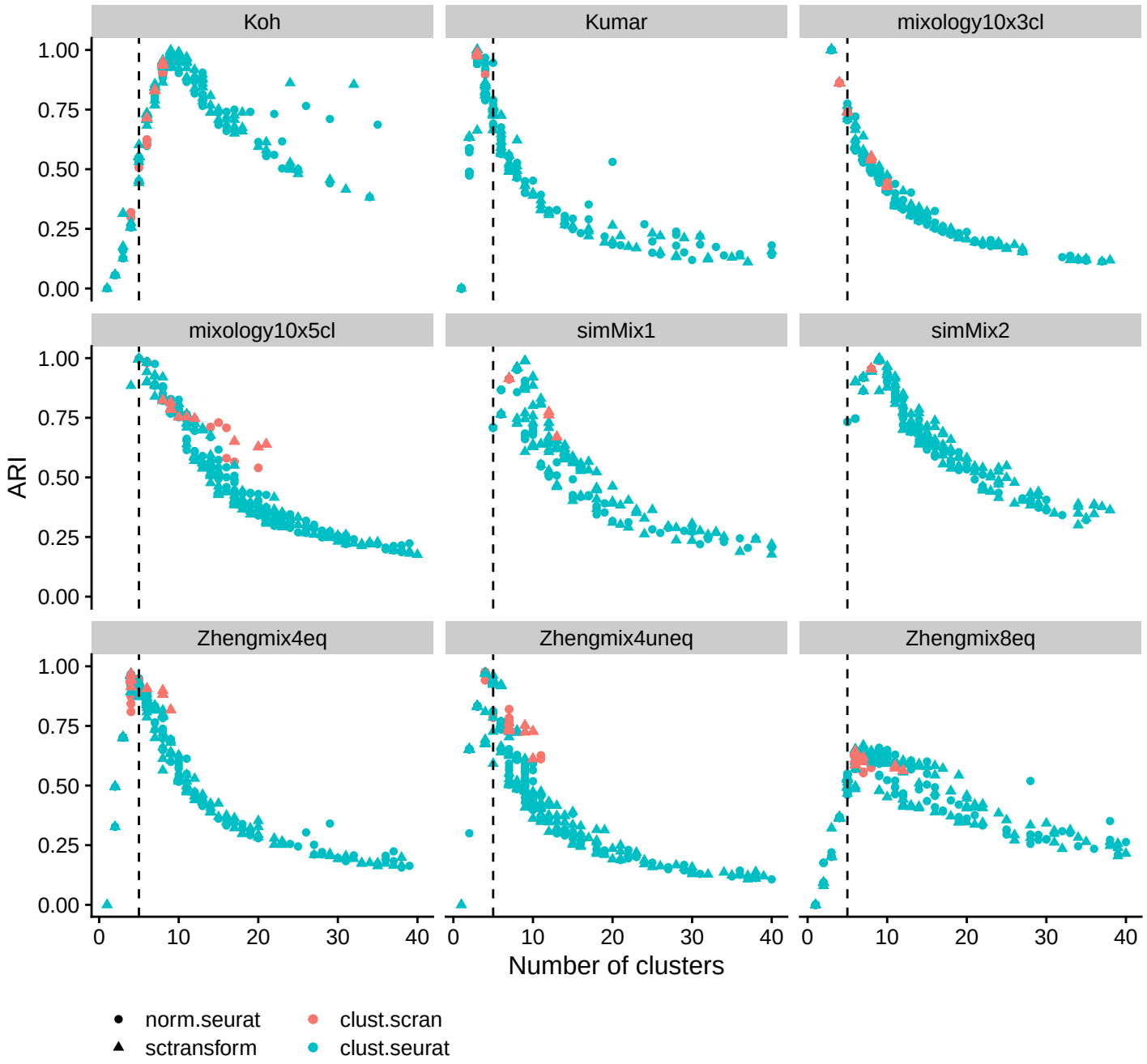
Pierre-Luc Germain
Anthony Sonrel
Mark D. Robinson

27 Februar, 2020

Contents

Supplementary Figure 1

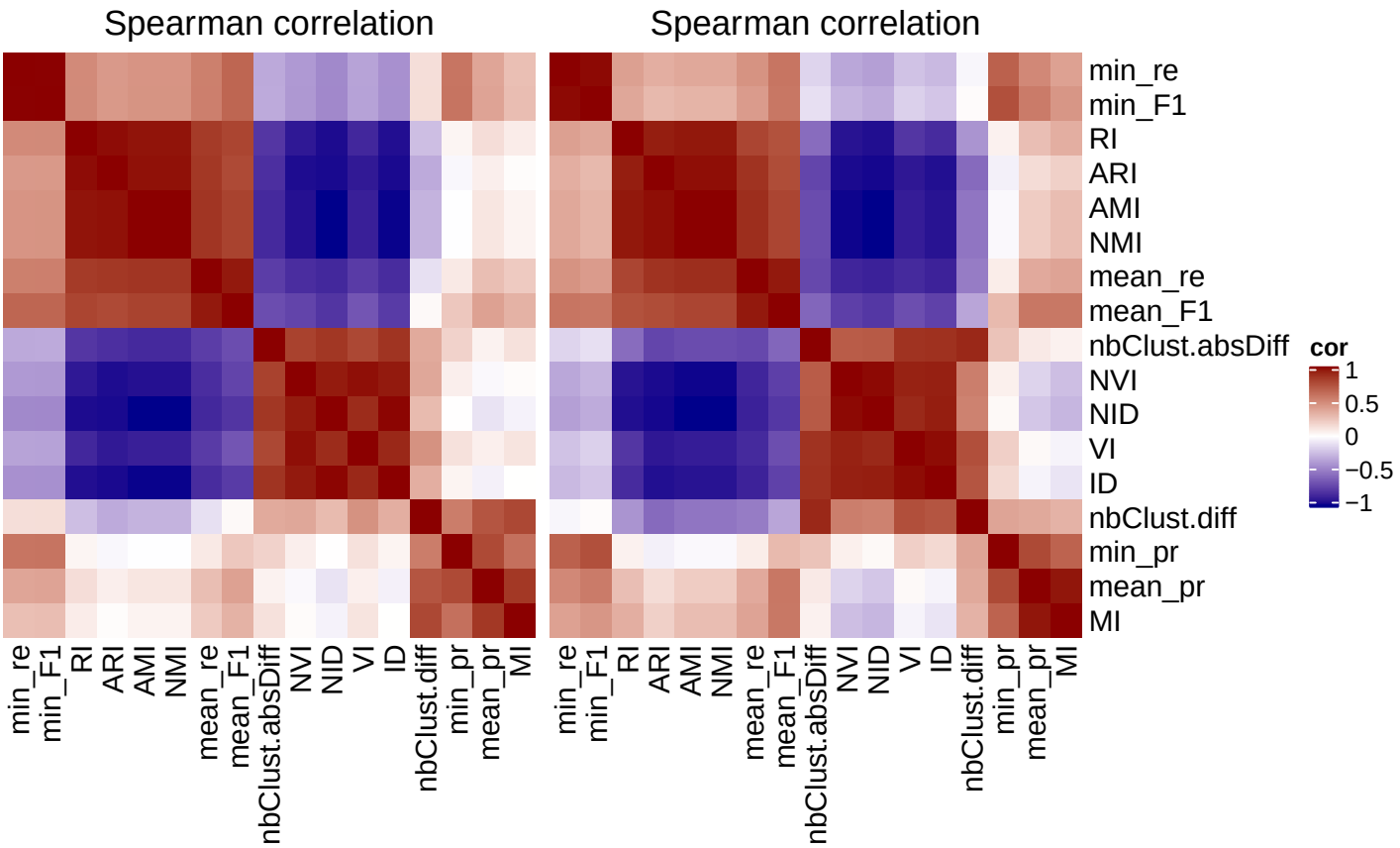
```
## Warning: Removed 169 rows containing missing values (geom_point).
```



Supplementary Figure 1

The number of clusters called has a much bigger impact on the Adjusted Rand Index (ARI) than differences between methods. The dashed line indicates the true number of clusters.

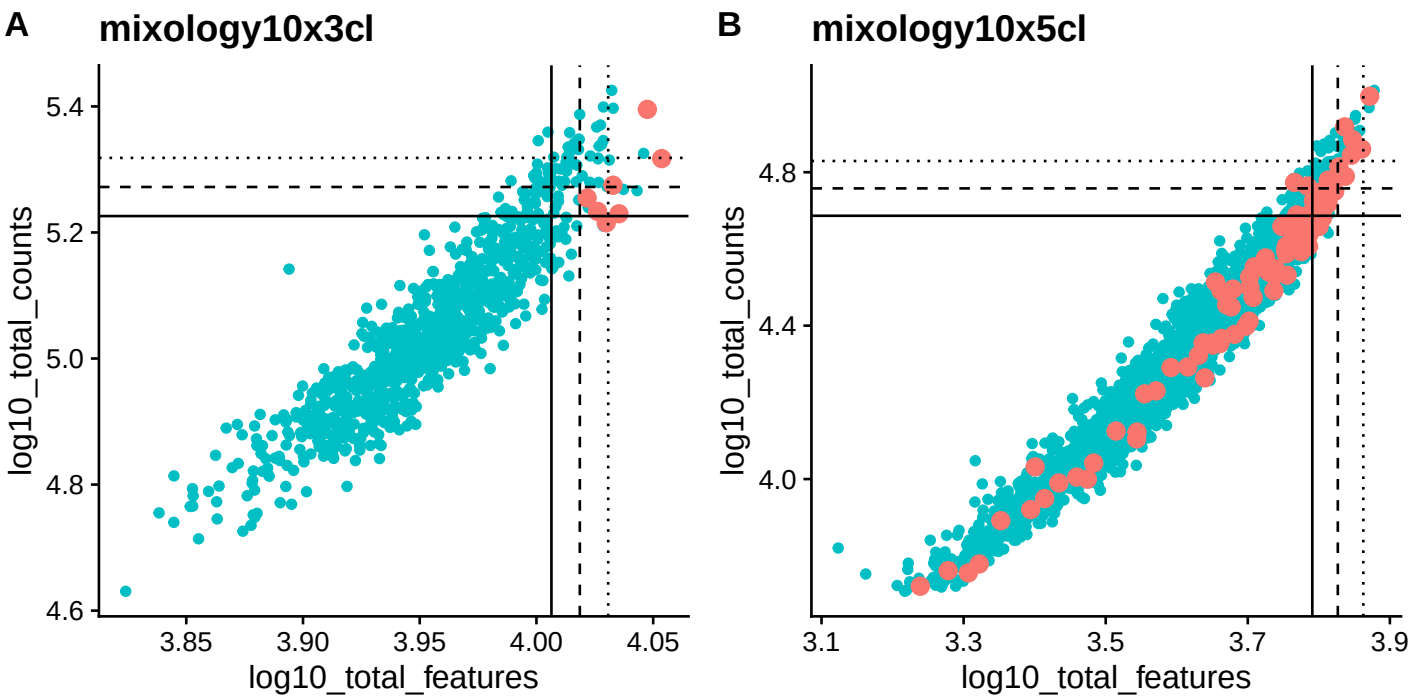
Supplementary Figure 2



Supplementary Figure 2

Relationship of various metrics of clustering accuracy between each other and with variations in the number of clusters called (`nbClust.diff` and `nbClust.absDiff`). Correlations were calculated for each dataset separately across various clustering runs and averaged (the `mixology10x3cl` dataset was excluded due to insufficient variation among the results). Information distance metrics (ID, NID, VI, NVI) are highly correlated with the absolute difference between the true and called number of clusters, while the Adjusted Rand Index (ARI) and similar metrics were strongly anticorrelated to it. Precision (mean_pr) and recall (mean_re) were slightly less correlated with discrepancies in the number of clusters. Mutual information (MI) was not at all correlated with the absolute difference in number of clusters (`nbClust.absDiff`), but positively correlated with the difference (`nbClust.diff`), i.e. favouring clusterings calling a higher number of clusters. We therefore recommend using complementary metrics such as ARI and MI, and potentially mean F1 per subpopulation.

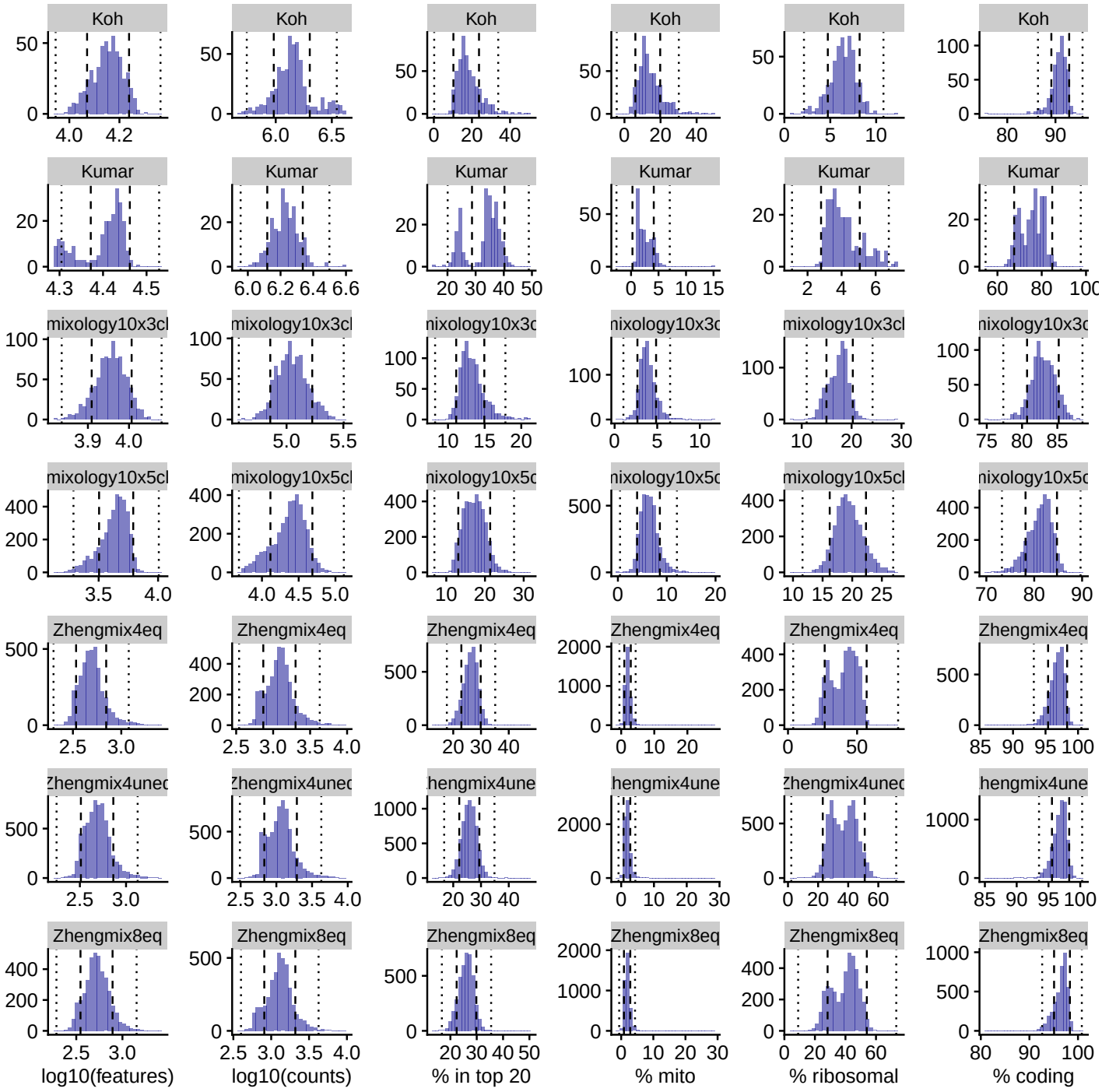
Supplementary Figure 3



Supplementary Figure 3

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.

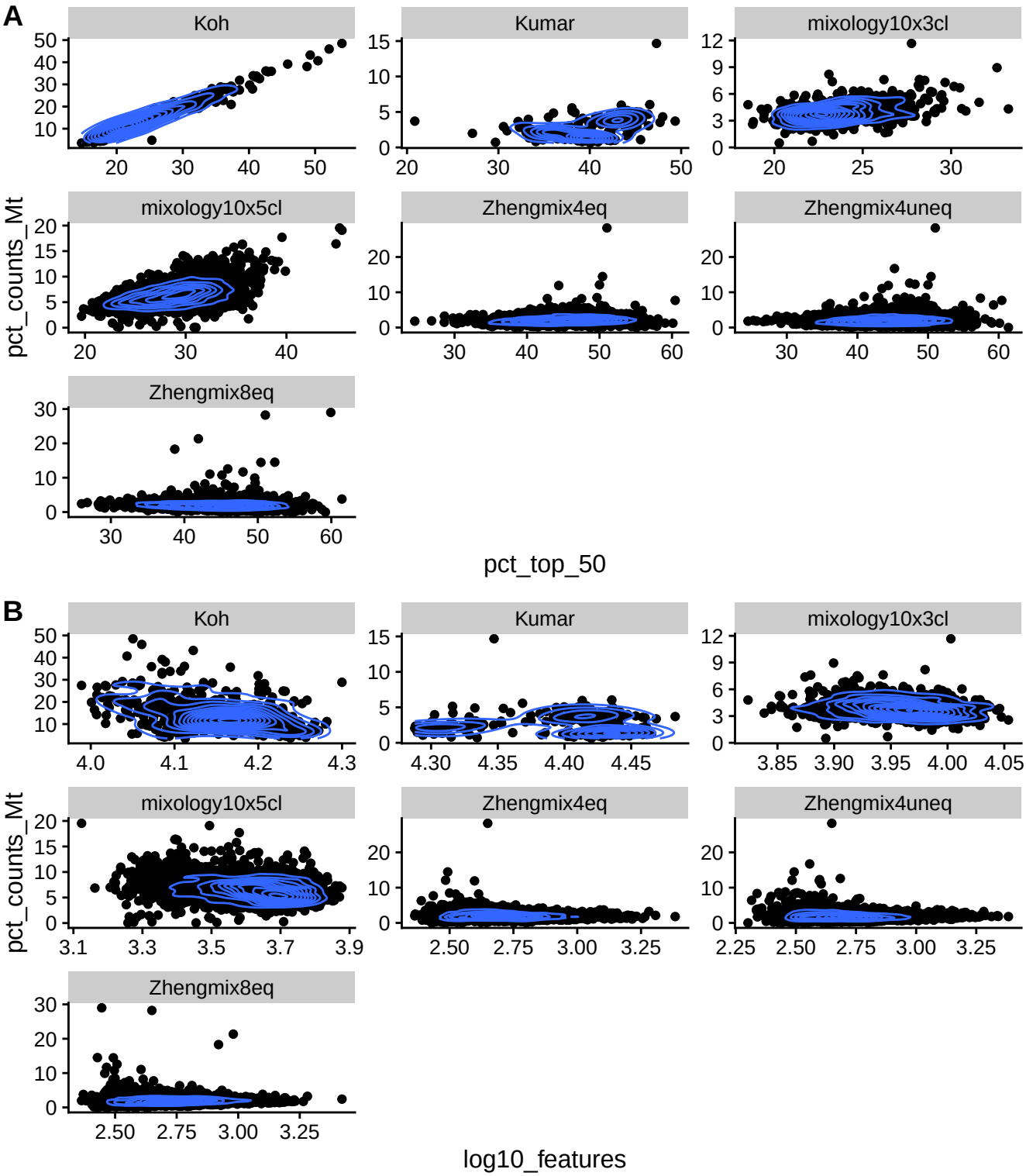
Supplementary Figure 4



Supplementary Figure 4

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

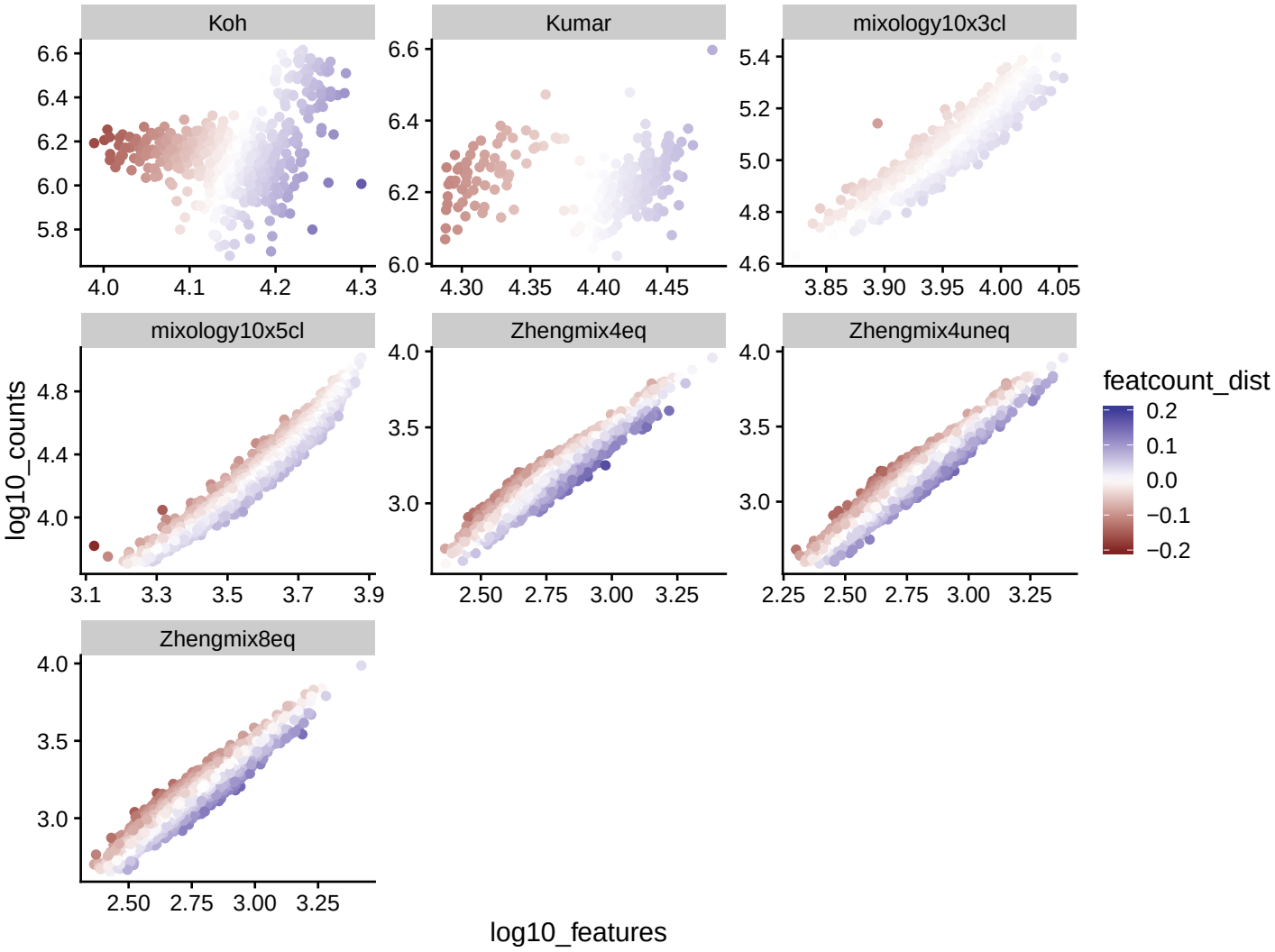
Supplementary Figure 5



Supplementary Figure 5

Relationship between selected cell-level QC metrics.

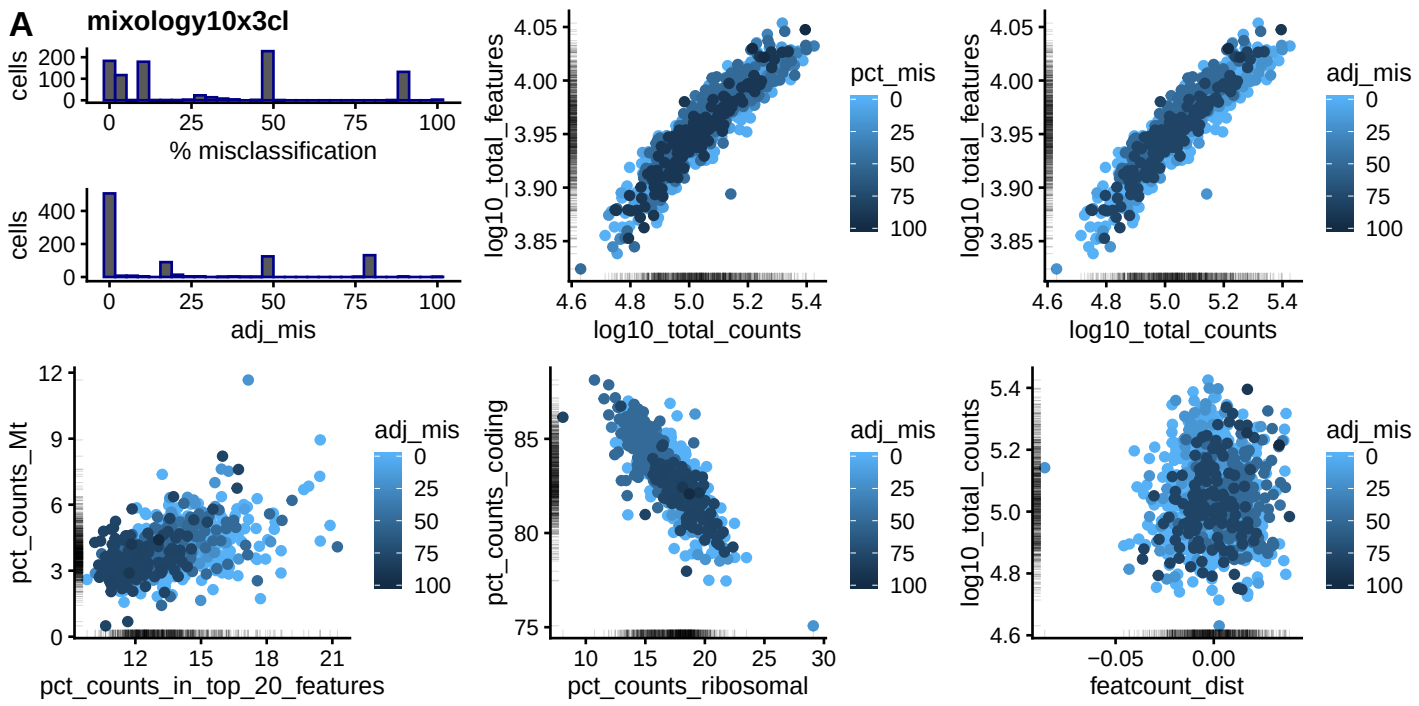
Supplementary Figure 6

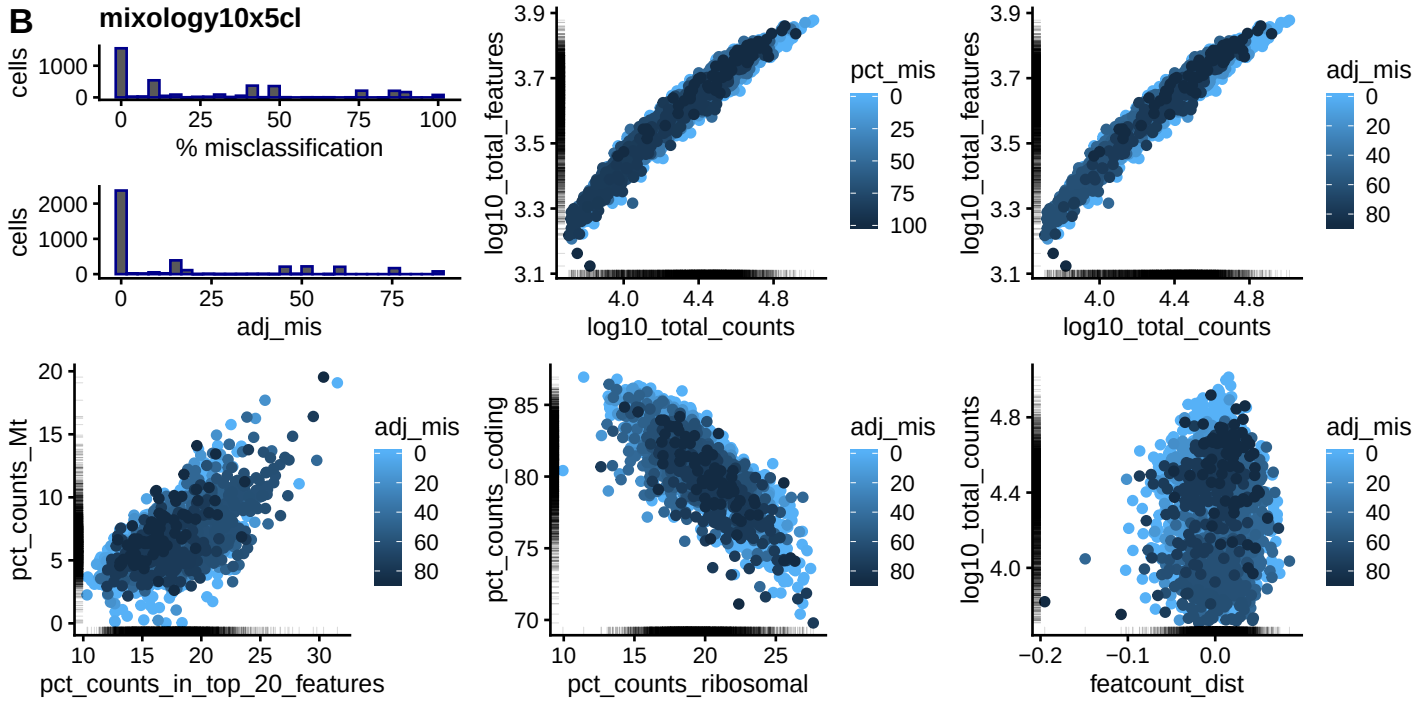


Supplementary Figure 6

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 7

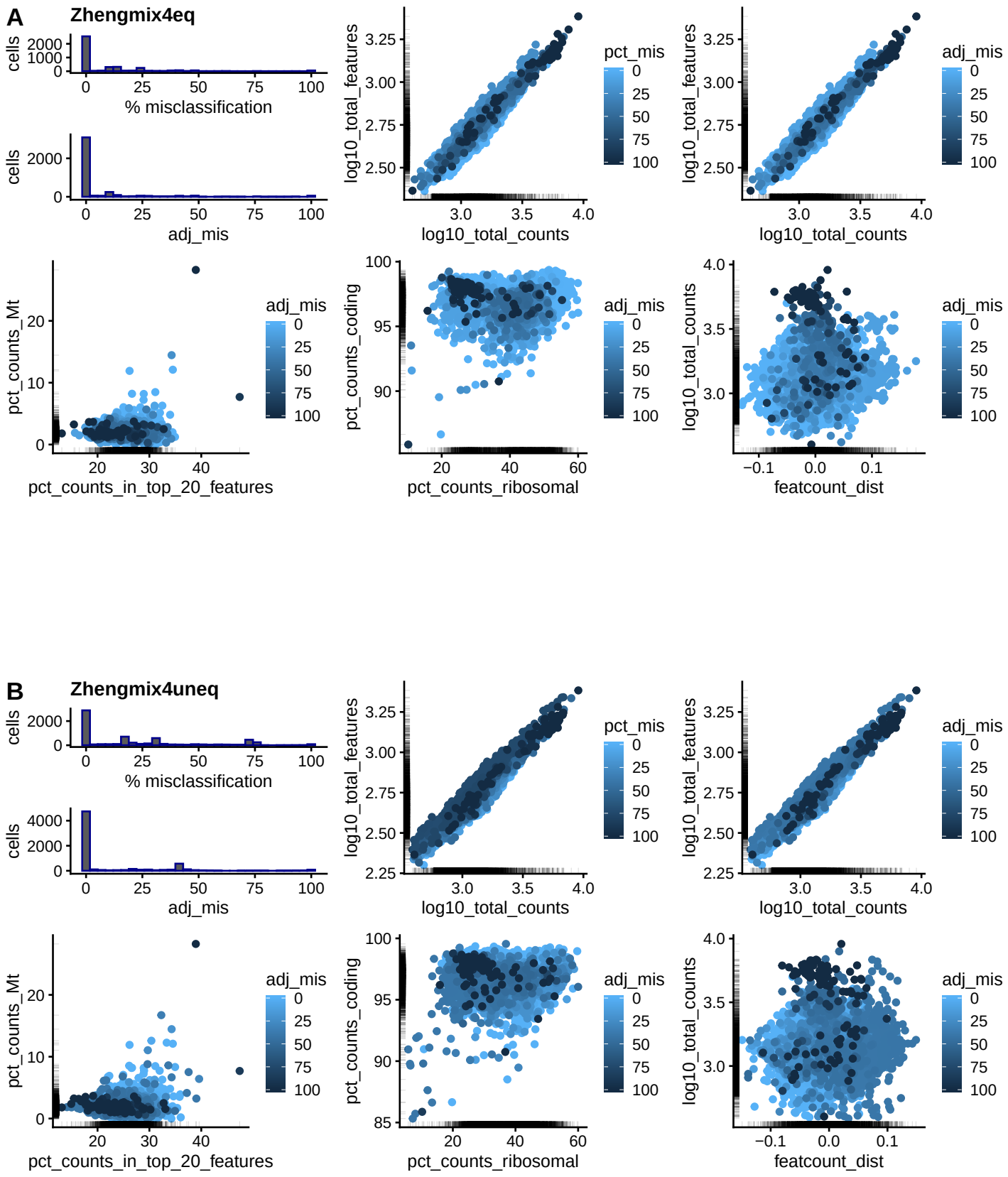




Supplementary Figure 7

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.

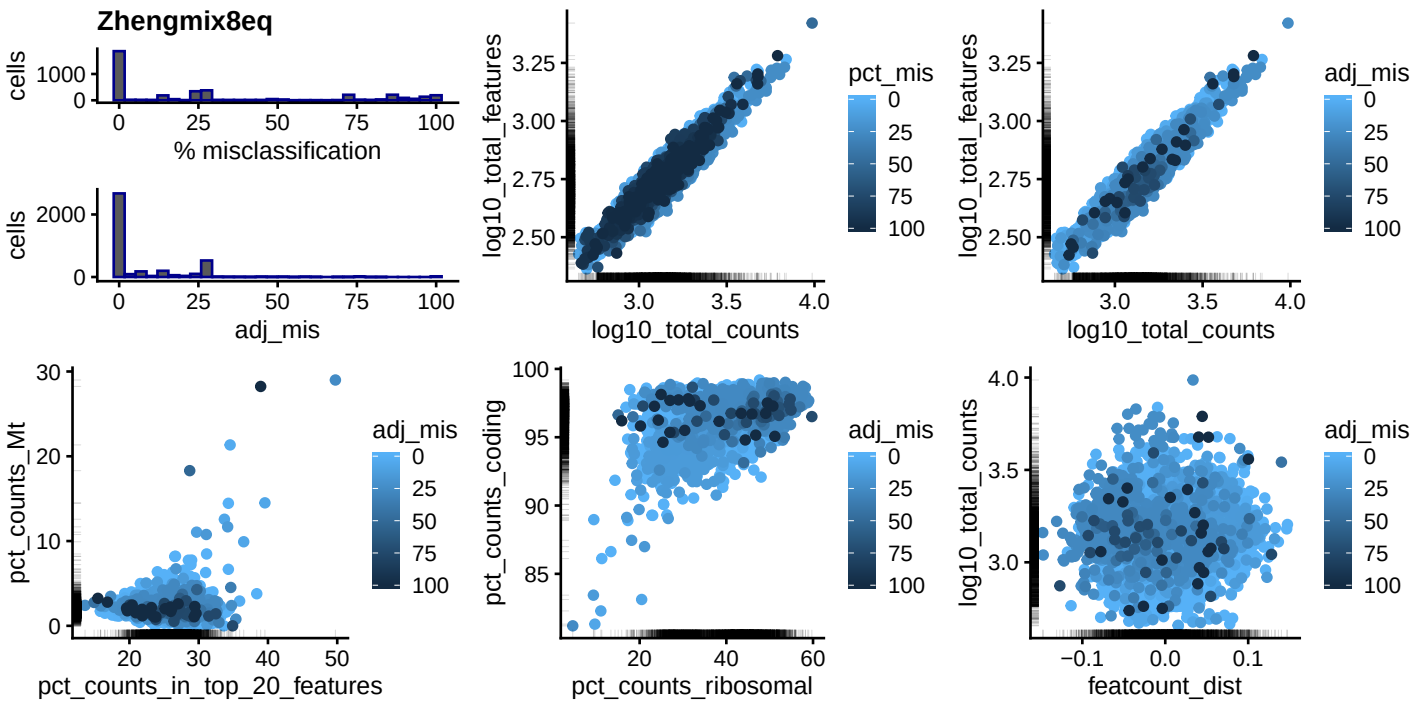
Supplementary Figure 8



Supplementary Figure 8

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 7 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.

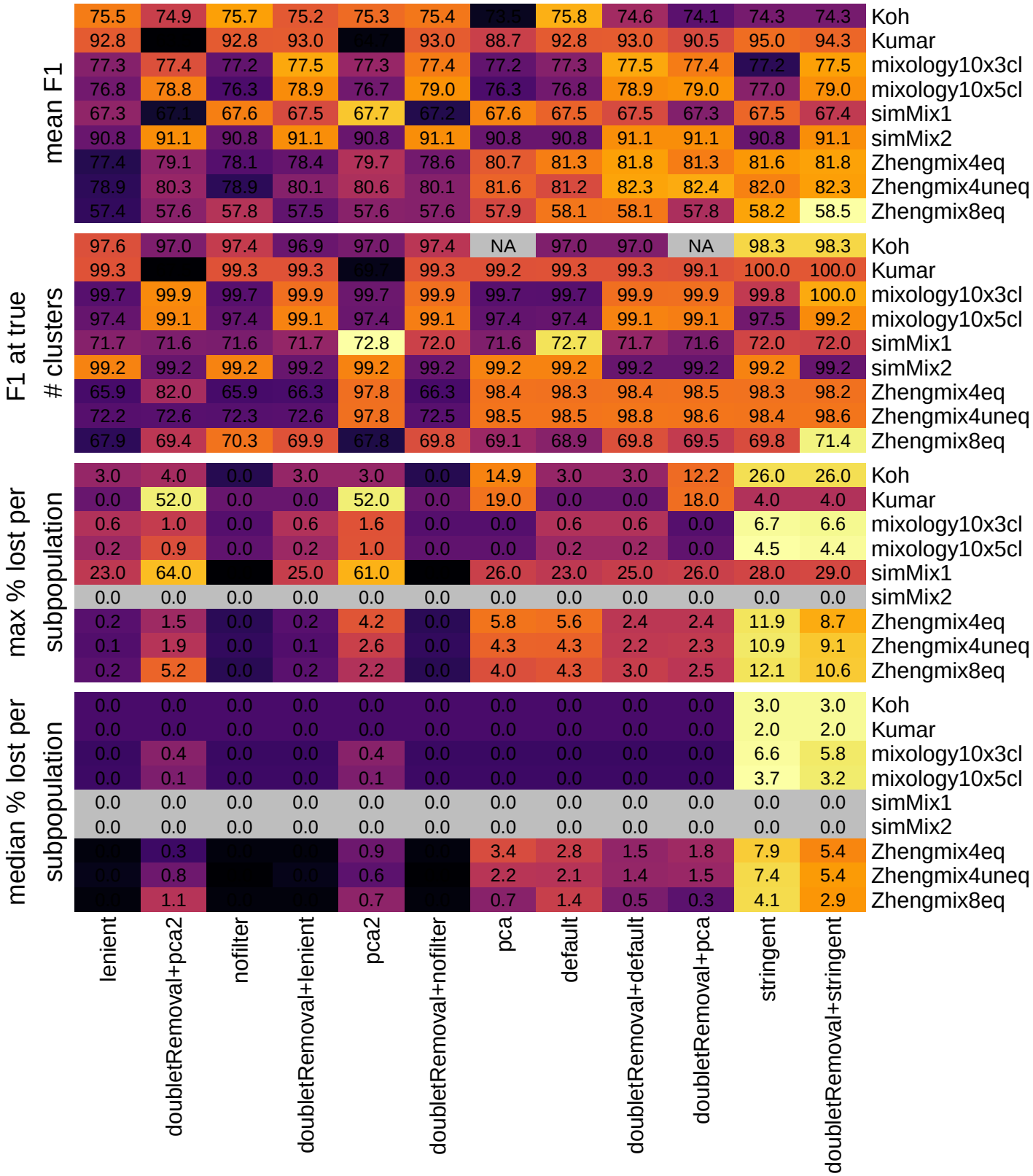
Supplementary Figure 9



Supplementary Figure 9

Relationship between various cellular properties and the frequency of cluster mis-assignment for the Zheng mixture of 8 cell types. See Supplementary Figure 7 for more information.

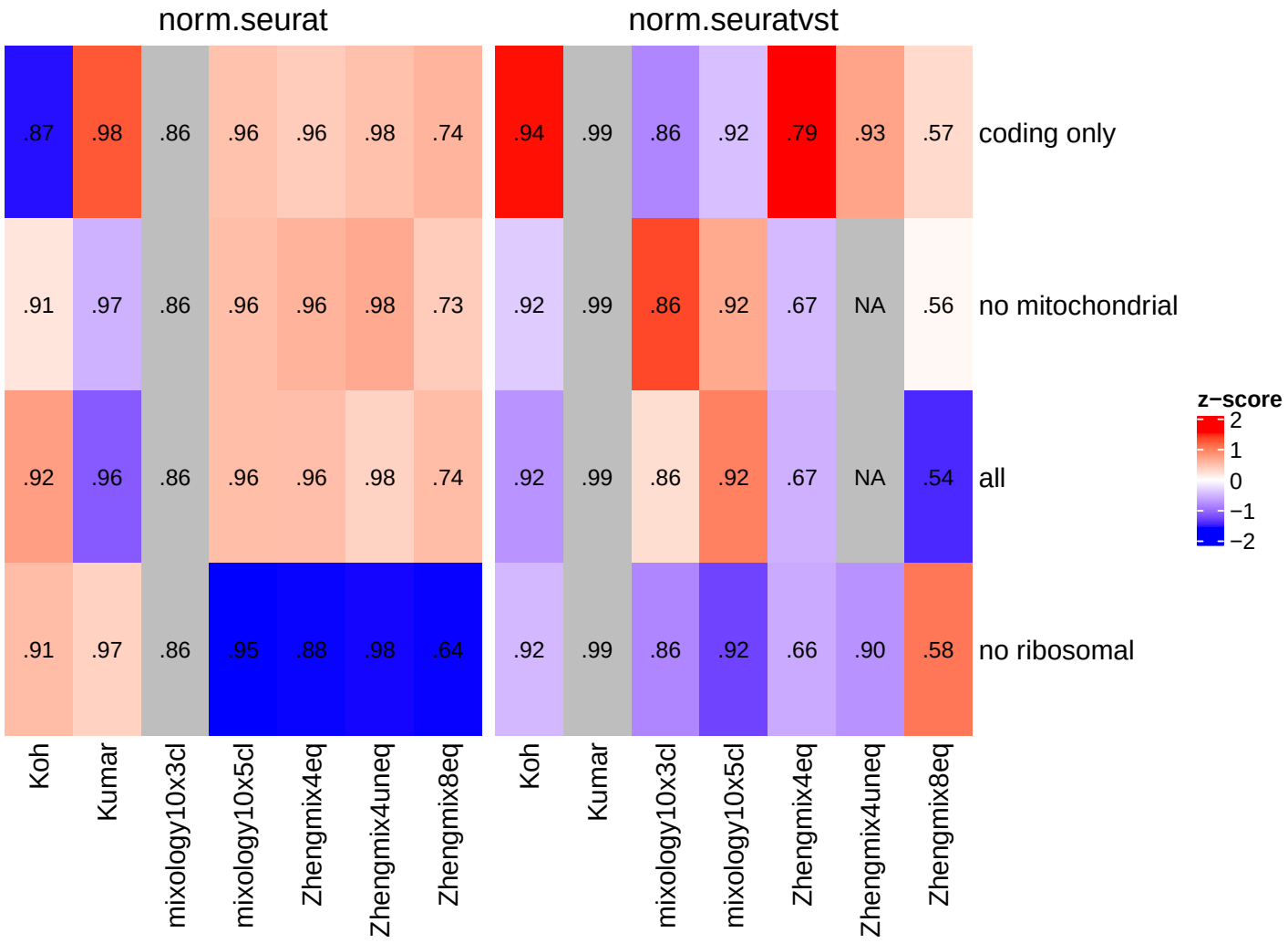
Supplementary Figure 10



Supplementary Figure 10

Mean clustering F1 score per subpopulation, mean F1 at true number of clusters, as well as maximum and median proportion of excluded cells per subpopulation across various filtering strategies. Doublet removal generally improves clustering accuracy with very mild exclusion rates, even in datasets that do not have heterotypic doublets. Stringent distribution-based filtering creates large cell type biases.

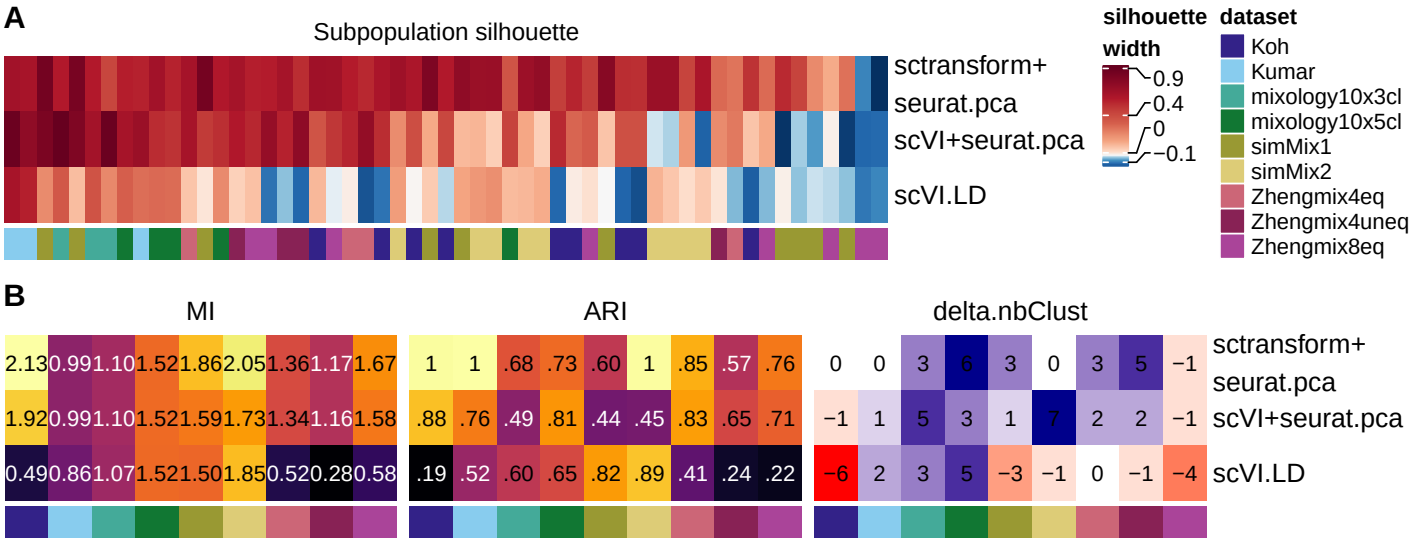
Supplementary Figure 11



Supplementary Figure 11

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

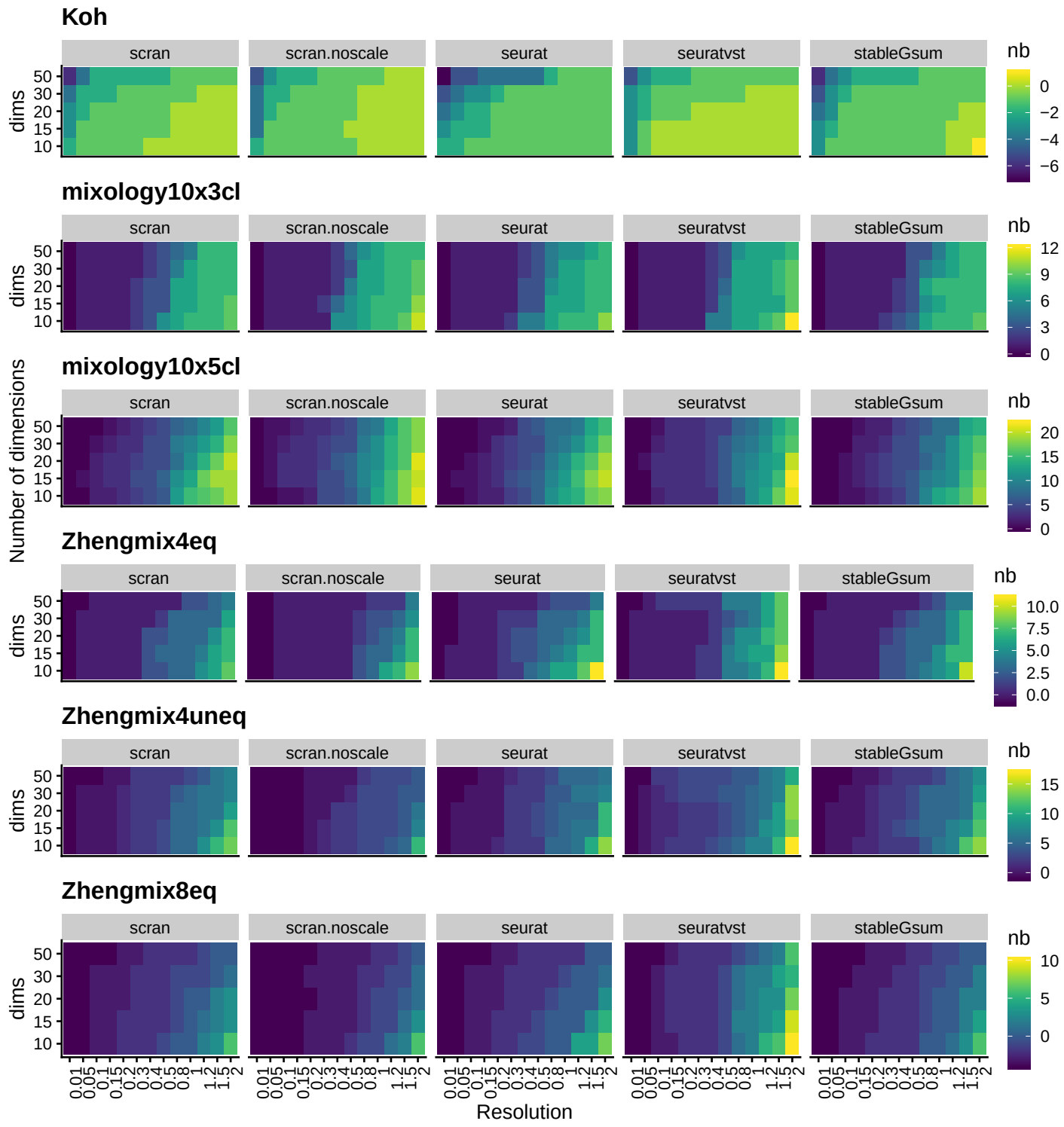
Supplementary Figure 12



Supplementary Figure 12

scVI evaluation. **A:** Average silhouette width per subpopulation using sctransform and standard Seurat PCA, scVI normalization followed by Seurat PCA, or scVI linear decoder (LD). **B:** Clustering accuracy across the same methods followed by Seurat clustering.

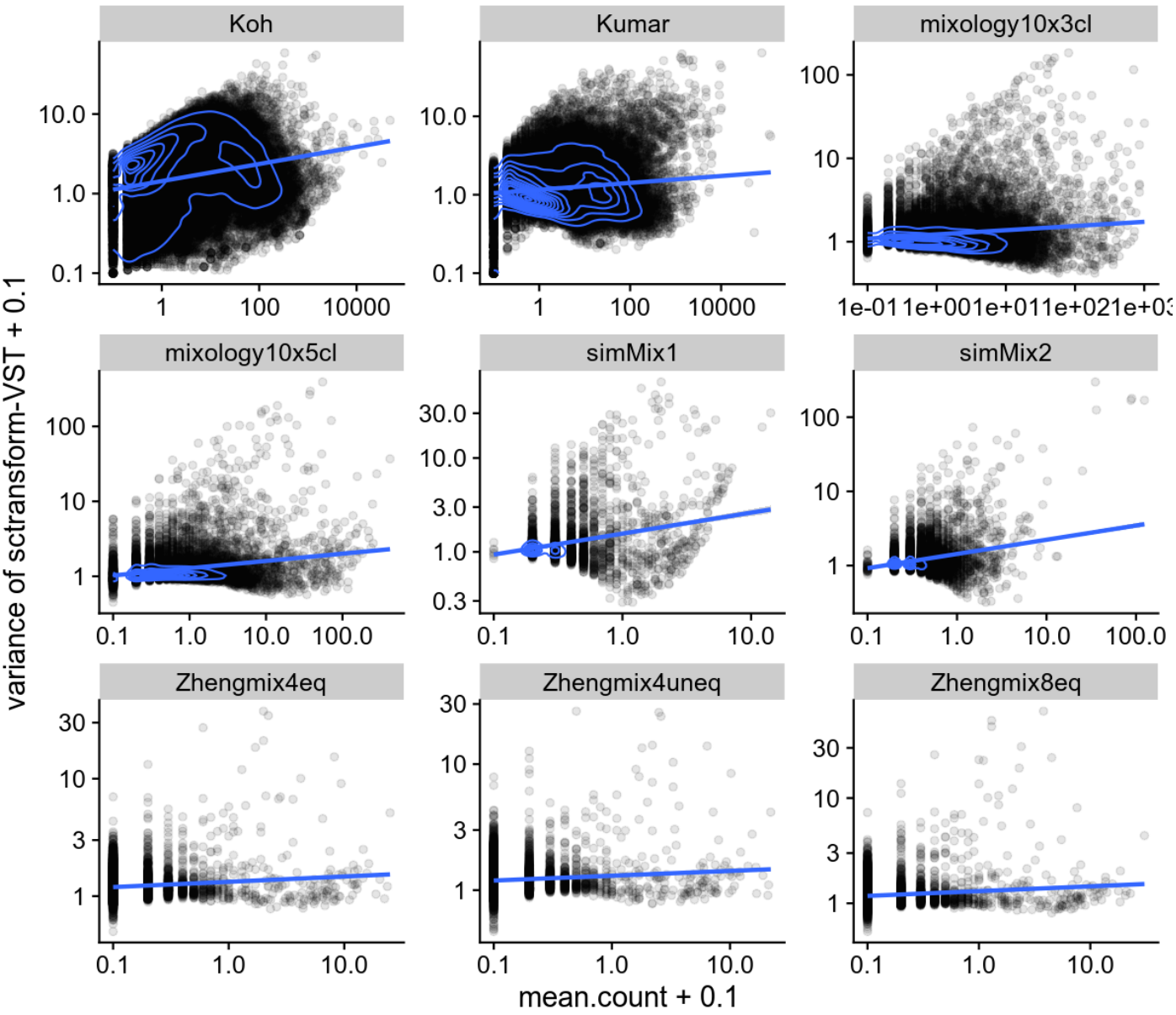
Supplementary Figure 13



Supplementary Figure 13

Mean difference between the number of detected clusters and the number of real subpopulations, depending on the normalization method, the resolution and the number of dimensions used. The Kumar dataset is not shown here due to a lack of variation in the number of clusters detected. A rough ANOVA on `nbClusters~dataset+norm+dims+resolution` suggests that seuratvst (sctransform) is associated with a higher number of clusters (p~0.002).

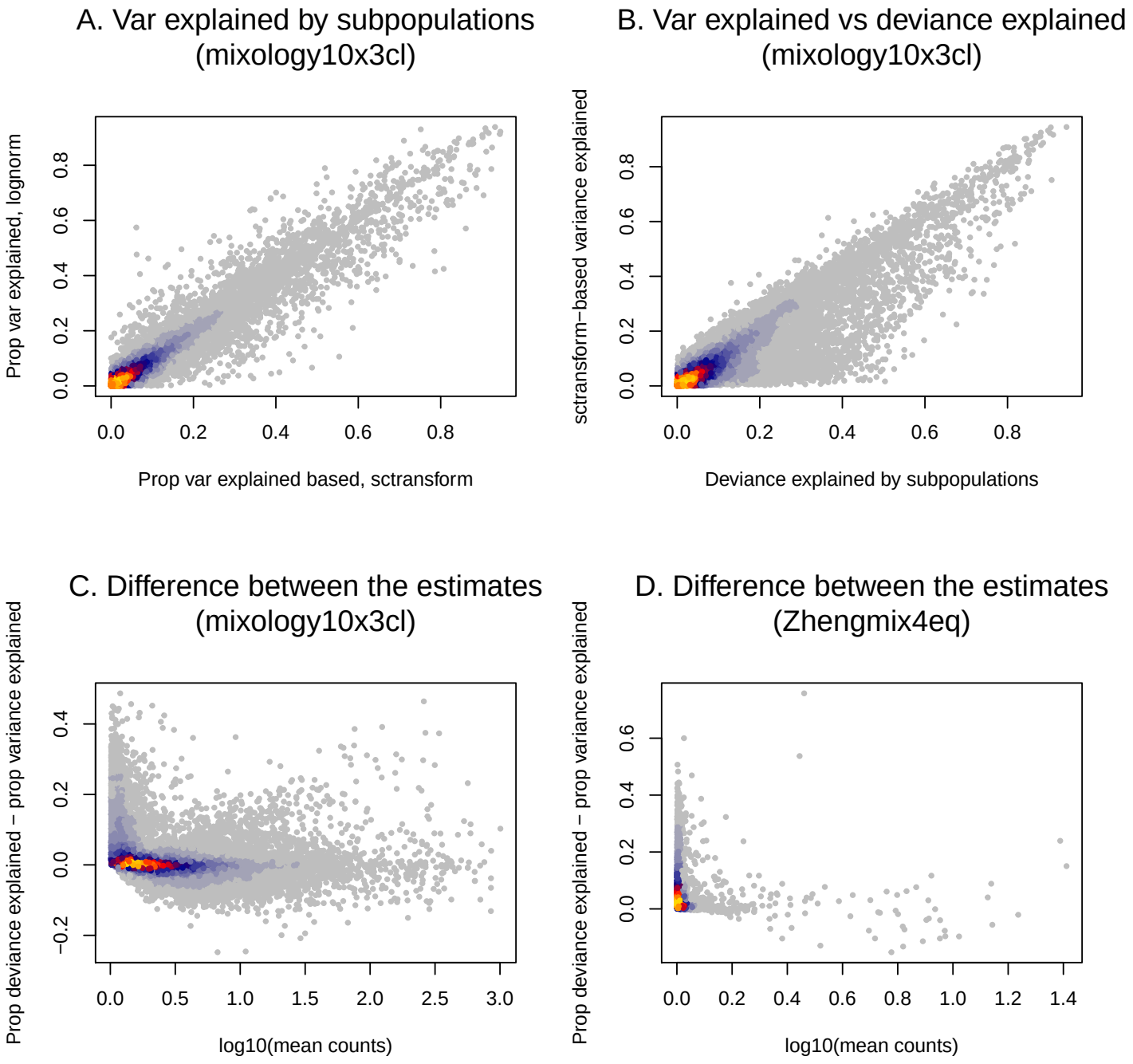
Supplementary Figure 14



Supplementary Figure 14

Relationship of the variance with mean count after `sctranform`'s variance stabilizing transformation.

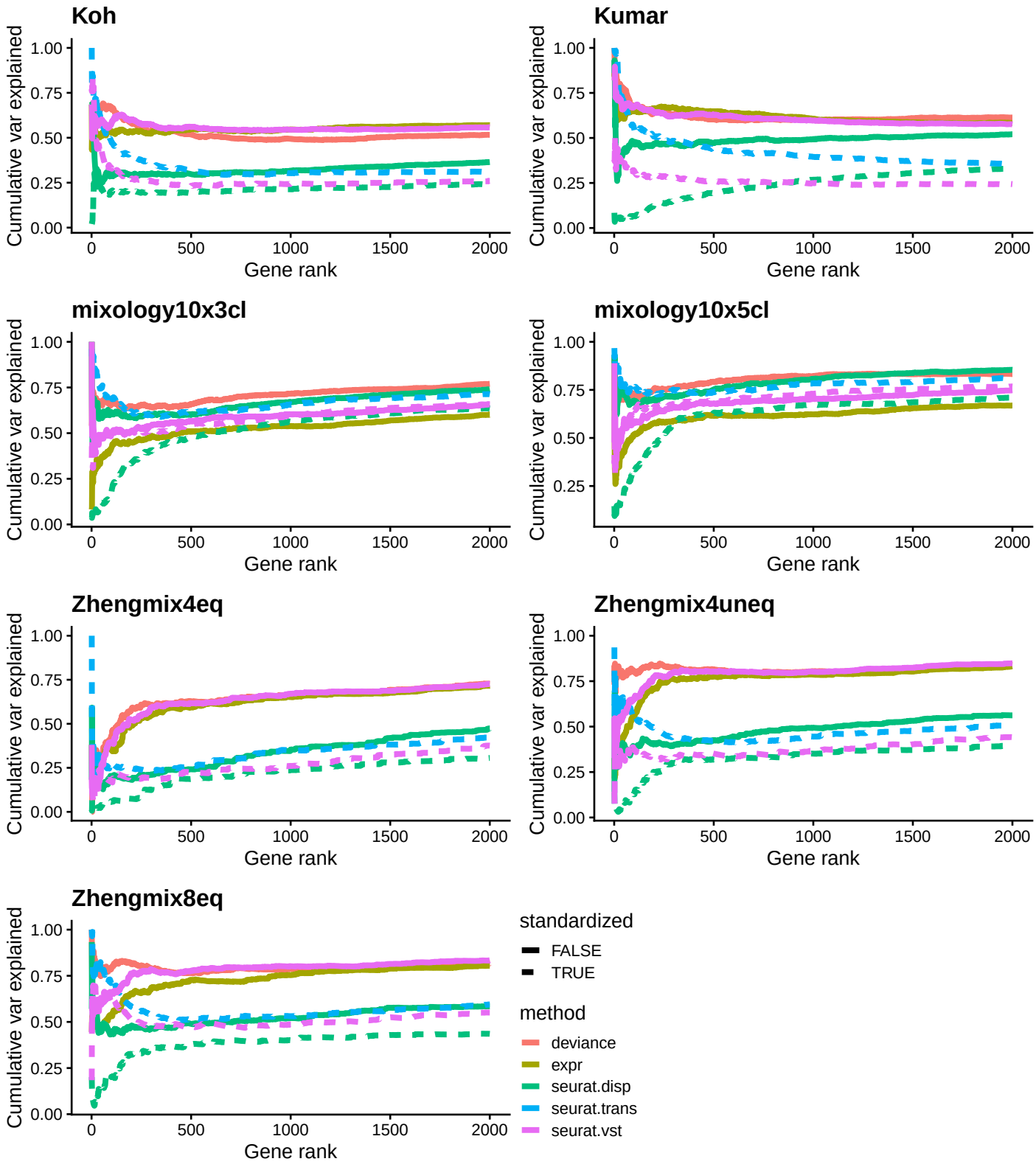
Supplementary Figure 15



Supplementary Figure 15

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.

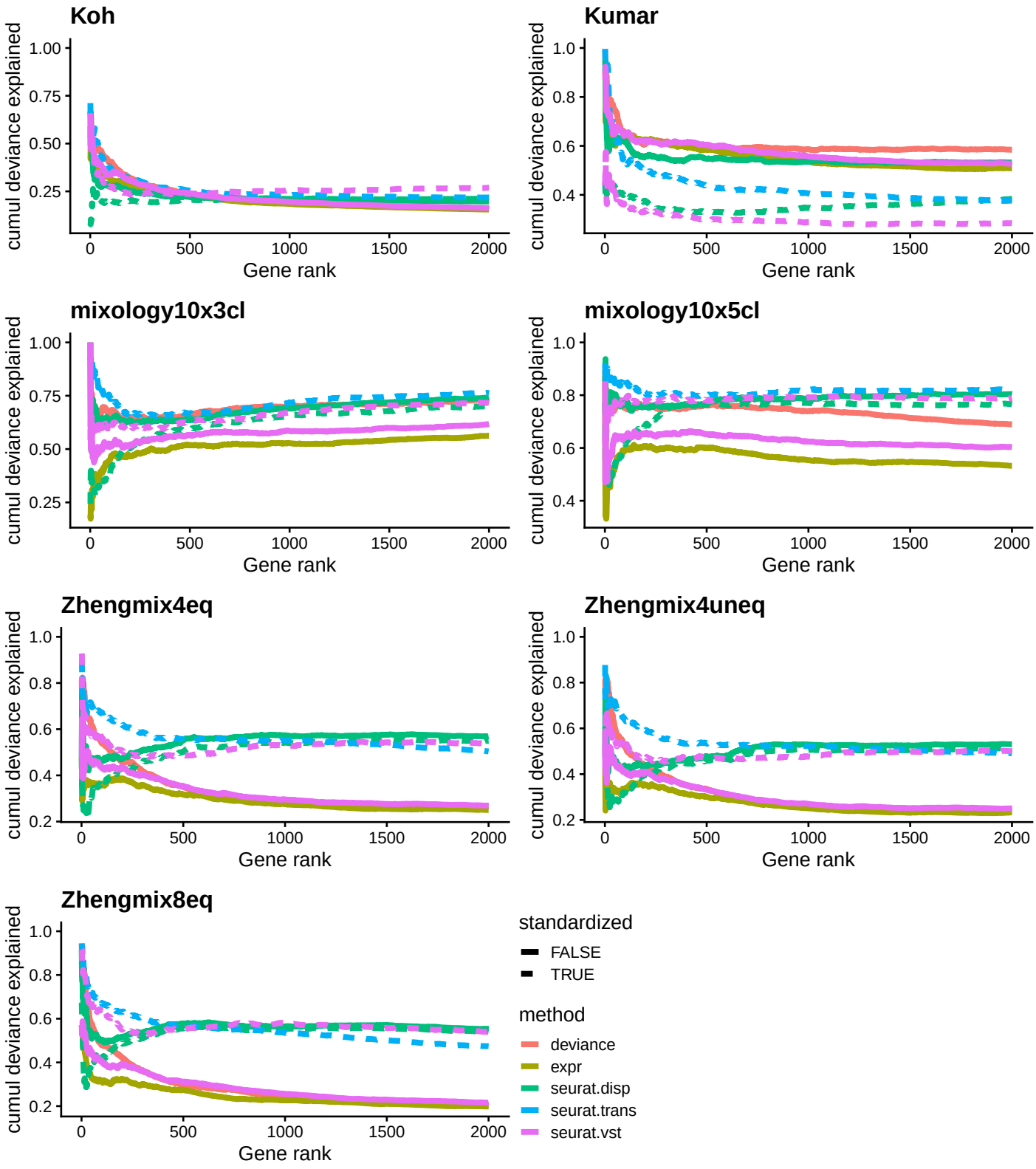
Supplementary Figure 16



Supplementary Figure 16

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.

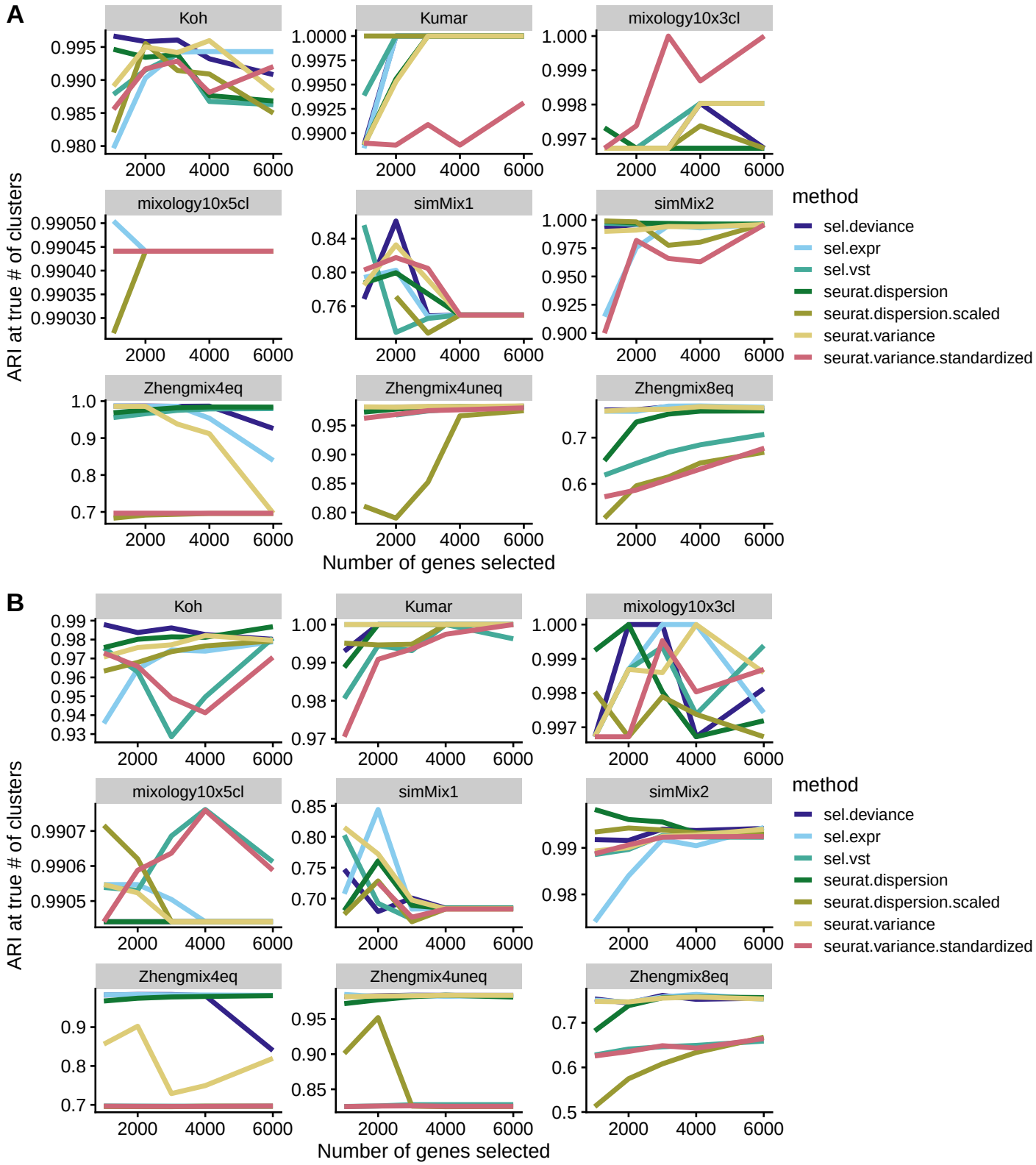
Supplementary Figure 17



Supplementary Figure 17

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 16, except using deviance explained.

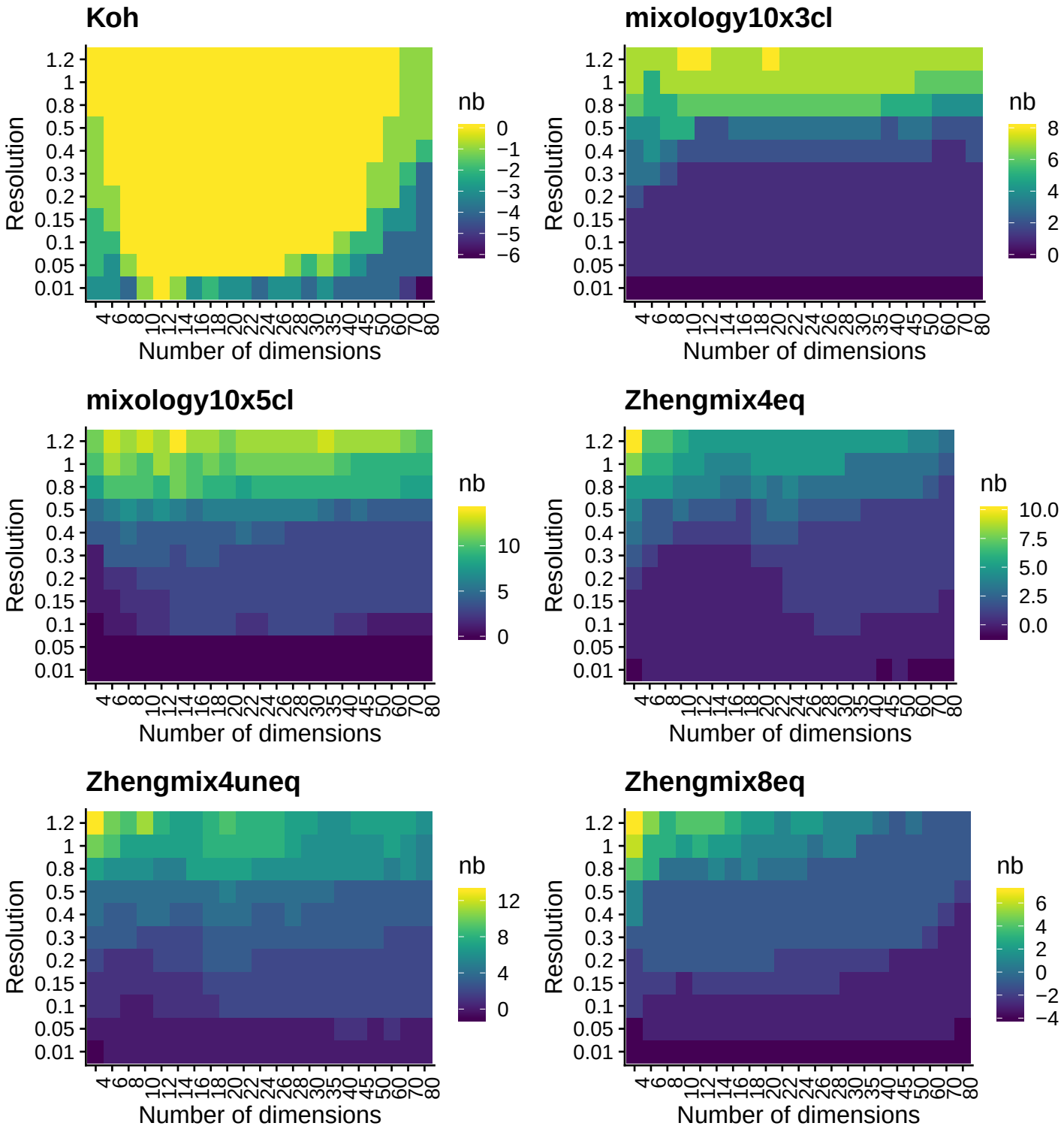
Supplementary Figure 18



Supplementary Figure 18

Clustering accuracy according to the number of genes selected using various ranking/selection methods. **A:** Based on sctransform, **B:** Based on standard Seurat normalization.

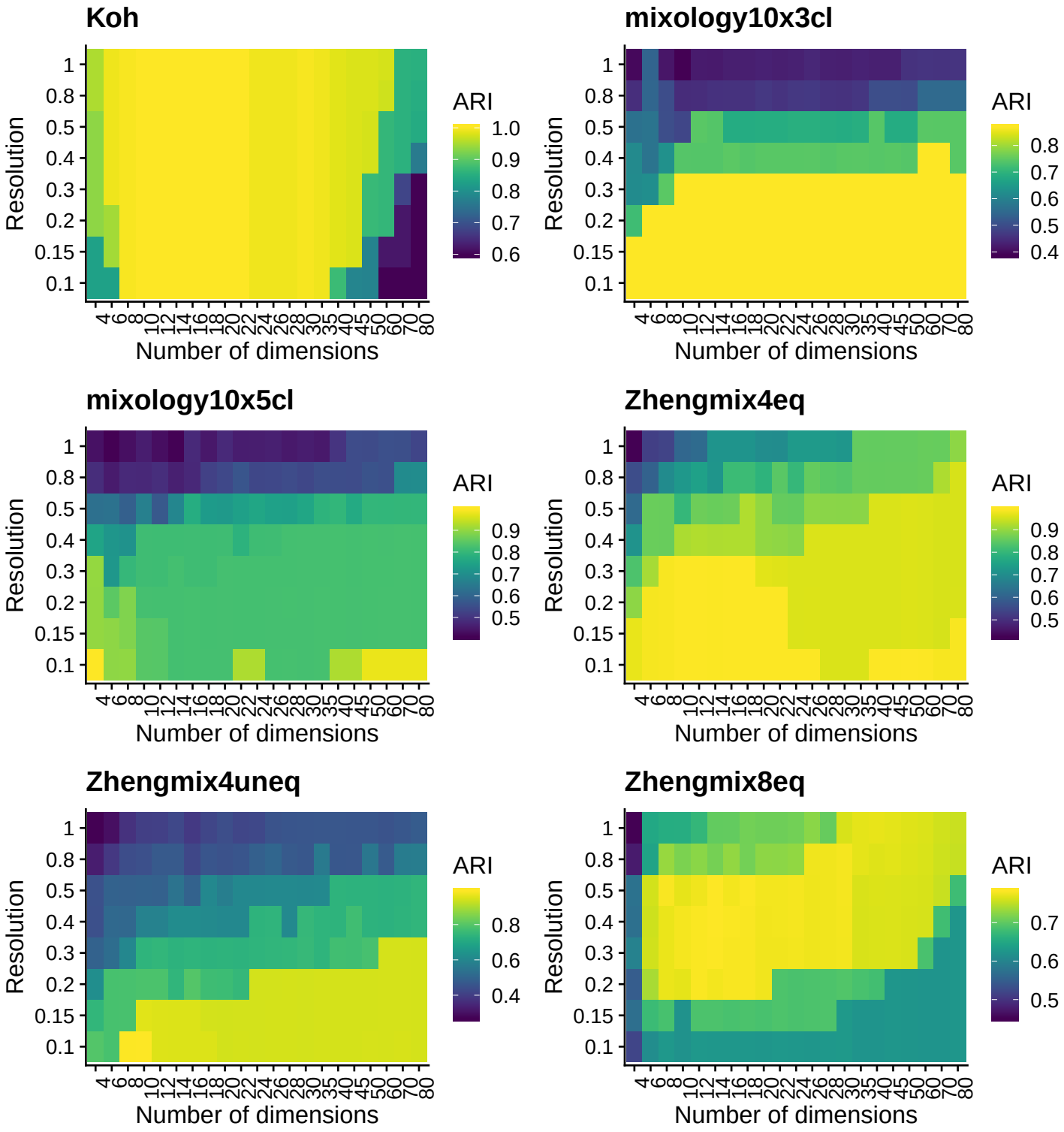
Supplementary Figure 19



Supplementary Figure 19

Mean difference between the number of detected clusters and the number of real subpopulations, depending on the resolution and number of dimensions used. Based on sctransform and seurat PCA. Increasing the number of dimensions tends to decrease the number of identified clusters, especially at resolutions around the default value.

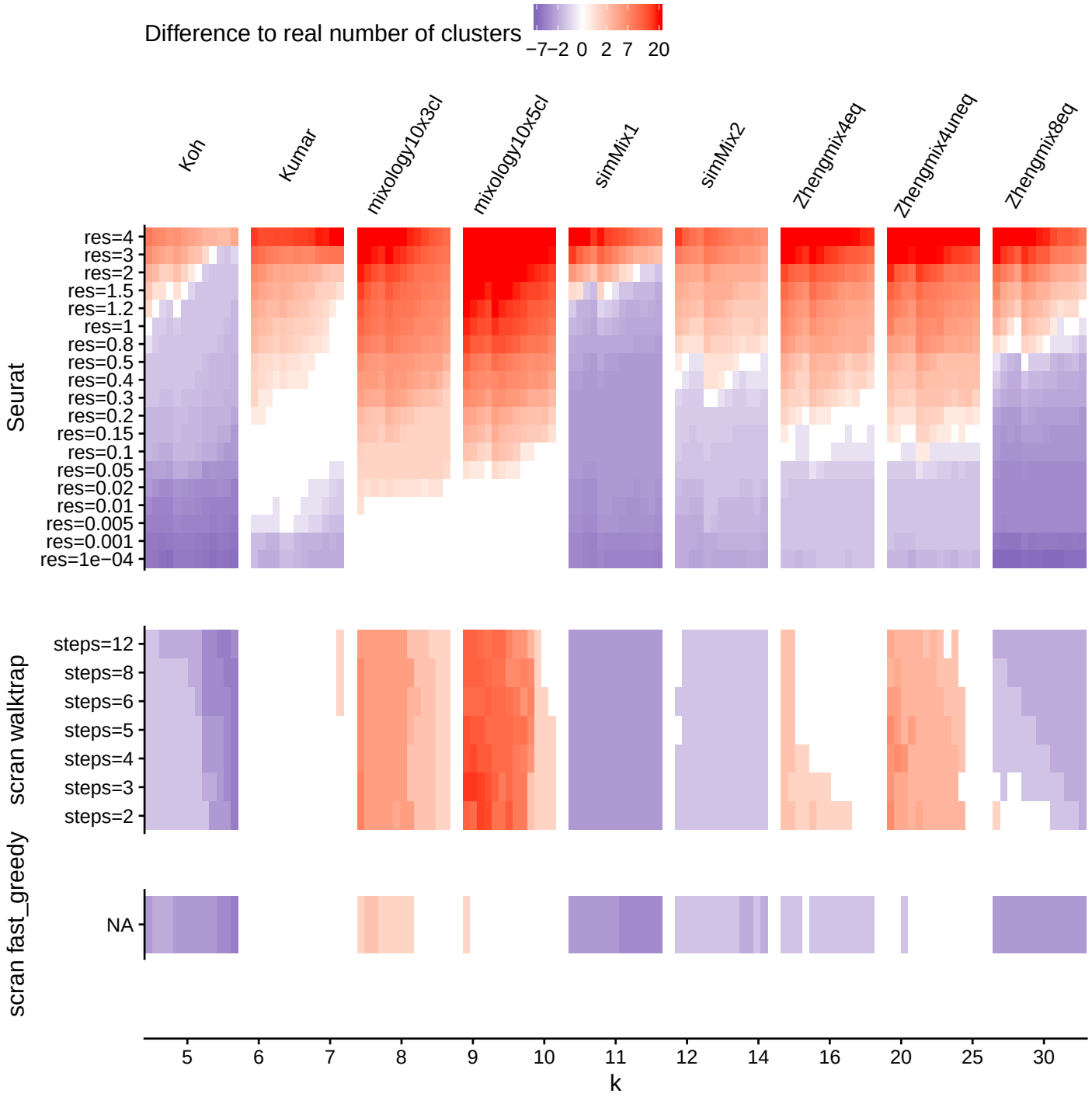
Supplementary Figure 20



Supplementary Figure 20

Adjusted Rand Index of clustering depending on the resolution and number of dimensions used. Based on sctransform and seurat PCA.

Supplementary Figure 21



Supplementary Figure 21

Difference between the number of detected clusters and the number of real subpopulations according to different clustering paramters.