

A rank-based marker selection method for high throughput scRNA-seq data: Additional file 1, supplementary figures

Anna C. Gilbert

Alexander Vargo

June 20, 2020

This file contains figures to supplement the content in the main manuscript.

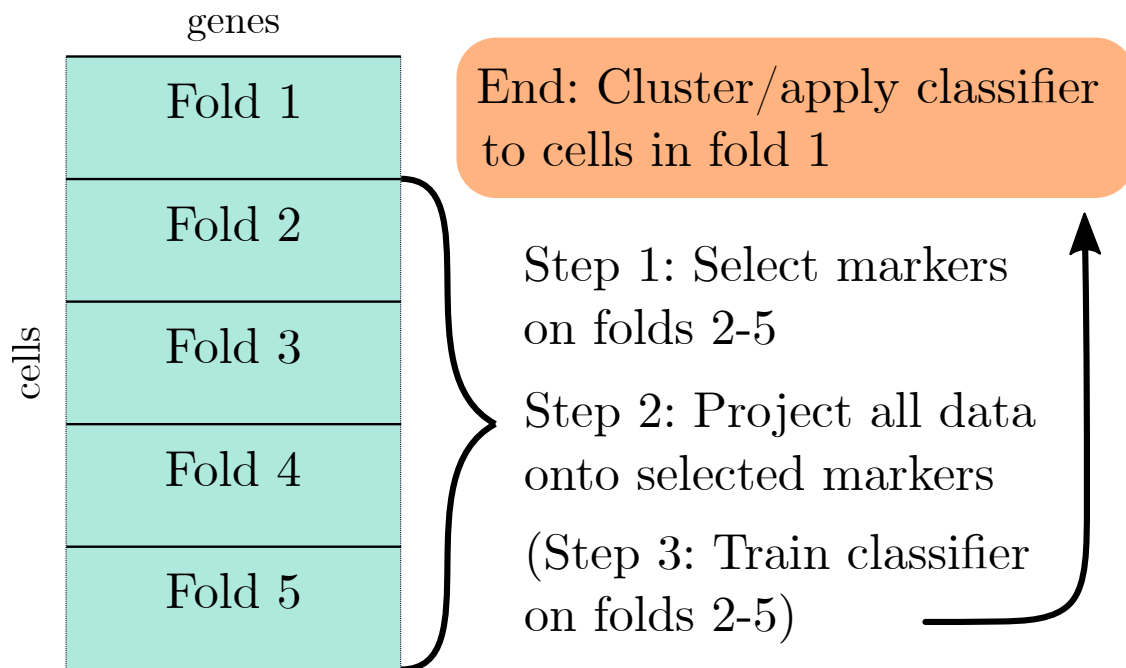


Figure 1: A visual description of 5 fold cross-validation

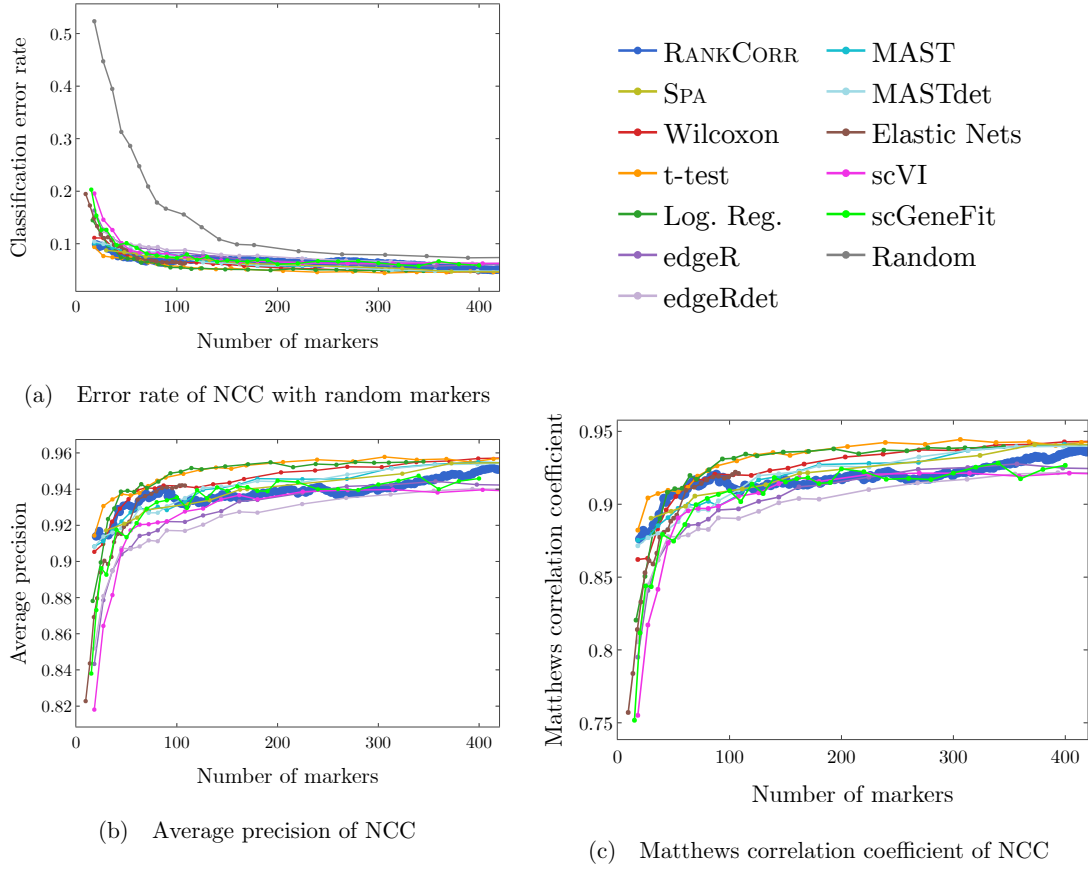


Figure 2: Supervised classification metrics for the ZEISEL data set using the nearest centroid classifier (NCC). Included is the performance of the scGeneFit method. (a) is the classification error rate including a curve corresponding to random marker selection (the curve that starts at around 50% error). (b) contains the precision and (c) contains the Matthews correlation coefficient. Data from random marker selection are not included in figures (b) and (c) for clarity. Note that the curves in (b) are similar in shape to the curves in (c); they are also similar in shape to the classification accuracy ($1 - \text{classification error rate}$ from Figure 4(a) of the main manuscript).

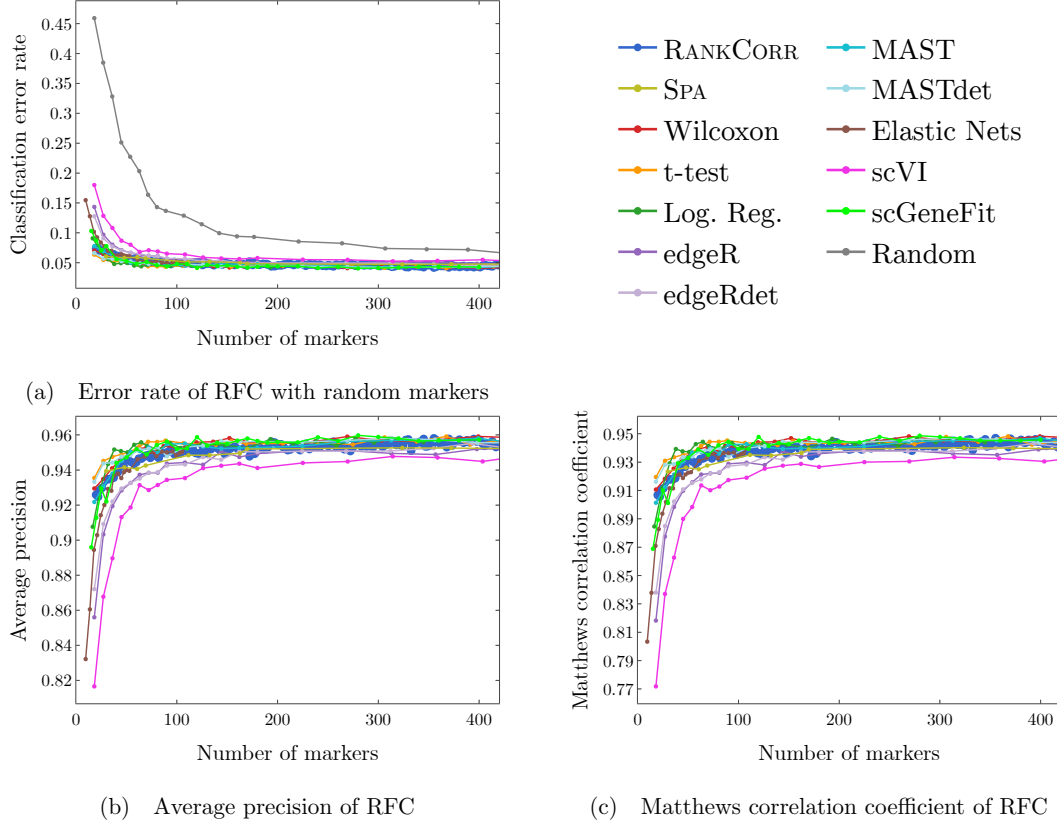


Figure 3: Supervised classification metrics for the ZEISEL data set using the random forests classifier (RFC). Included is the performance of the scGeneFit method. (a) is the classification error rate including a curve corresponding to random marker selection (the curve that starts at around 45% error). (b) contains the precision and (c) contains the Matthews correlation coefficient. Data from random marker selection are not included in figures (b) and (c) for clarity. Note that the curves in (b) are similar in shape to the curves in (c); they are also similar in shape to the classification accuracy ($1 - \text{classification error rate}$ from Figure 4(c) of the main manuscript).

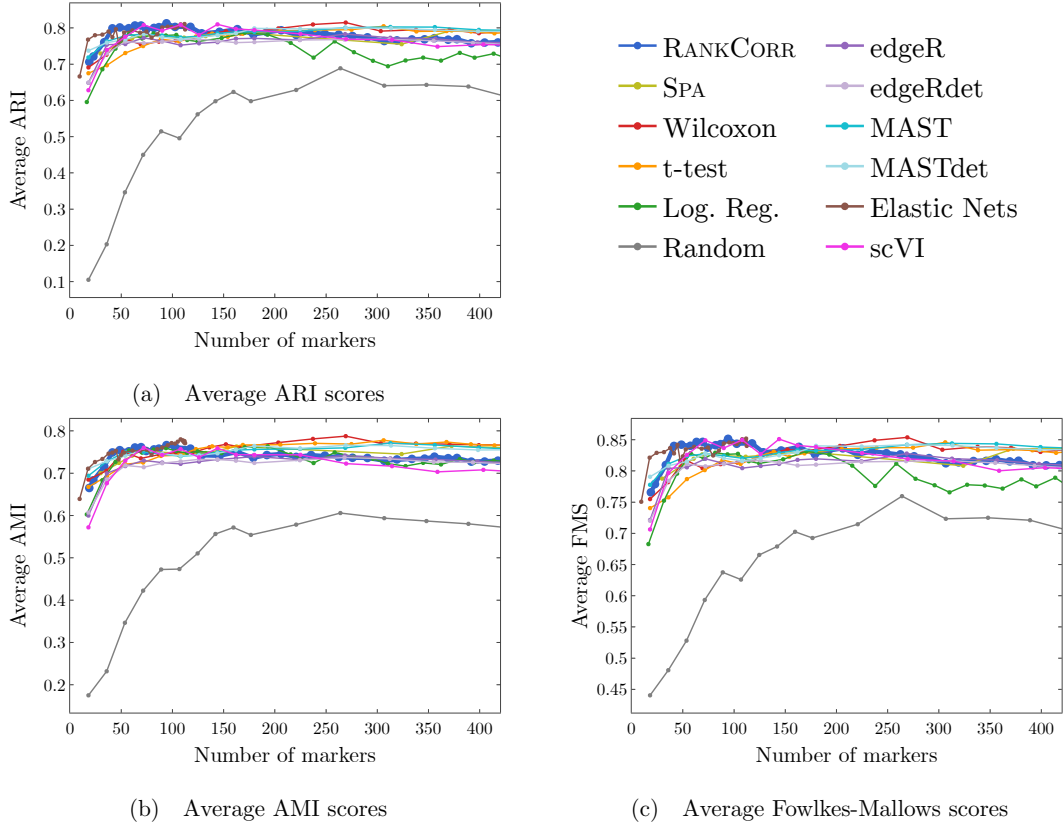


Figure 4: Unsupervised clustering metrics for the ZEISEL data set including data from random marker selection. The curve corresponding to random marker selection is shown in grey and is the lowest (worst) curve in all three plots. The ARI score is shown in (a), the AMI score is shown in (b), and the Fowlkes-Mallows score is shown in (c). The clustering is carried out using 5-fold cross validation and scores are averaged across folds.

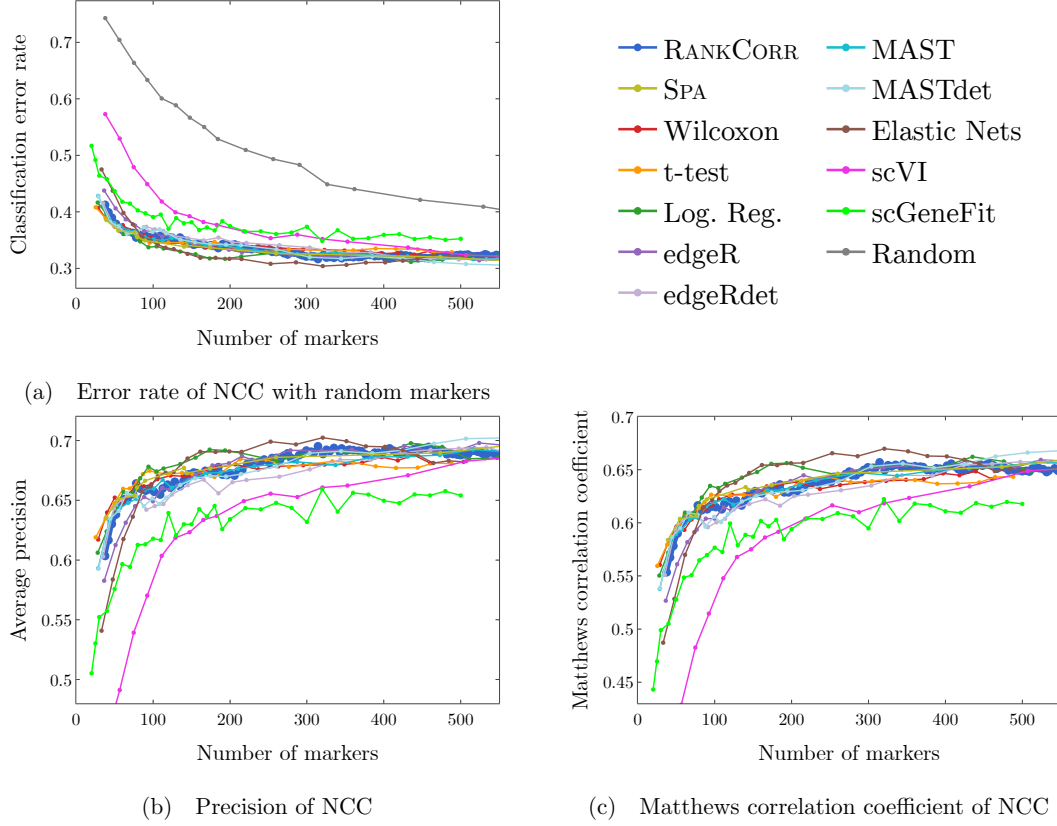


Figure 5: Supervised classification metrics for the PAUL data set using the nearest centroid classifier (NCC). Included is the performance of the scGeneFit method. (a) is the classification error rate including a curve corresponding to random marker selection (the curve that starts at around 75% error). (b) contains the precision and (c) contains the Matthews correlation coefficient. Data from random marker selection are not included in figures (b) and (c) for clarity. Note that the curves in (b) are similar in shape to the curves in (c); they are also similar in shape to the classification accuracy ($1 - \text{classification error rate}$ from Figure 6(a) of the main manuscript).

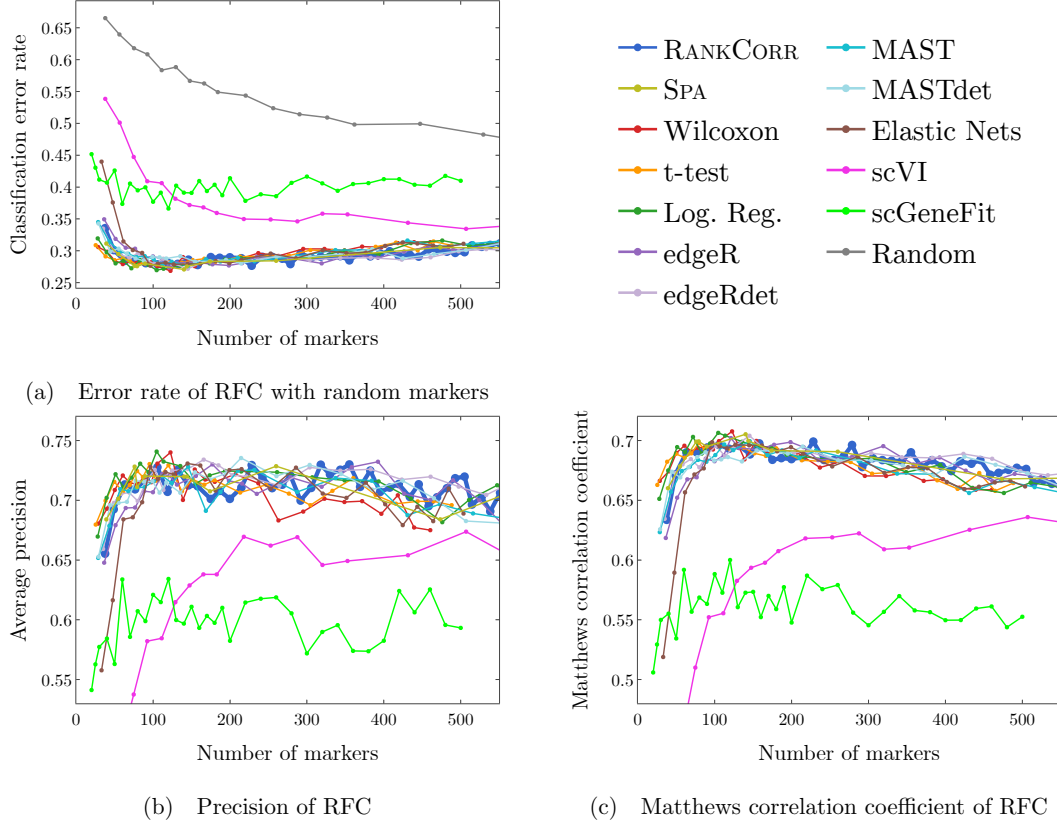


Figure 6: Supervised classification metrics for the PAUL data set using the random forests classifier (RFC). Included is the performance of the scGeneFit method. (a) is the classification error rate including a curve corresponding to random marker selection (the curve that starts at around 65% error). (b) contains the precision and (c) contains the Matthews correlation coefficient. Data from random marker selection are not included in figures (b) and (c) for clarity. Note that the curves in (b) are similar in shape to the curves in (c); they are also similar in shape to the classification accuracy ($1 - \text{classification error rate}$ from Figure 6(c) of the main manuscript).

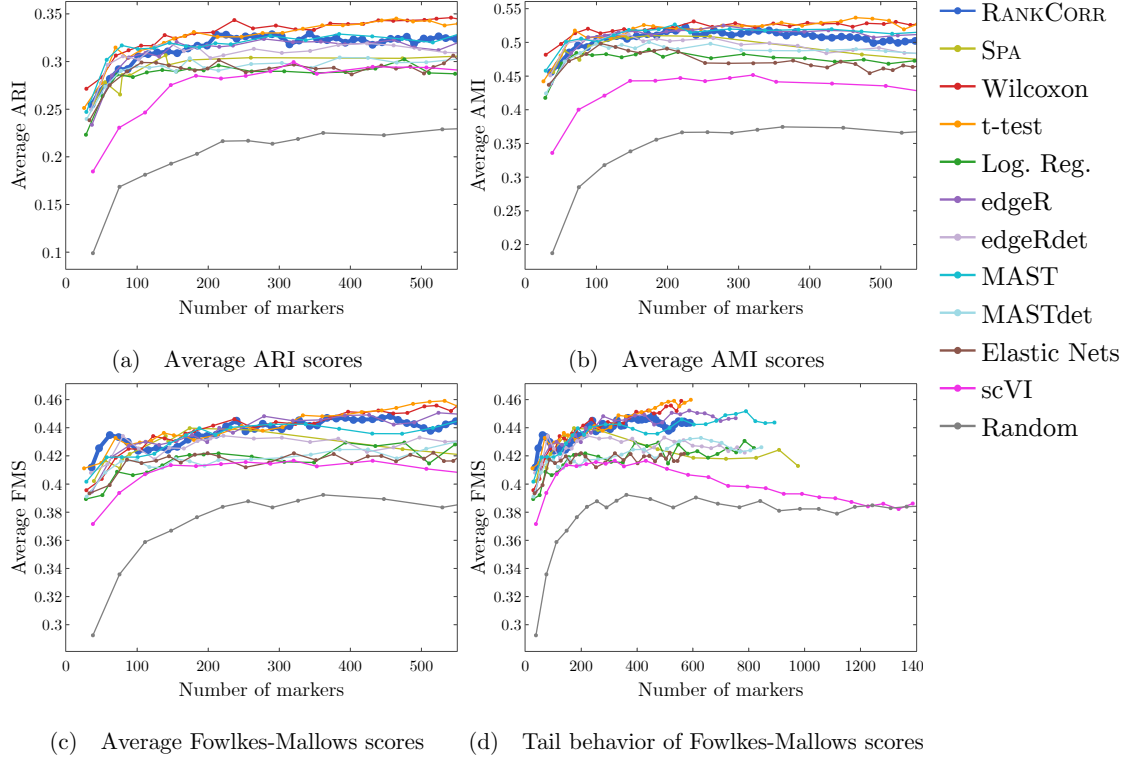


Figure 7: Unsupervised clustering metrics for the PAUL data set including data from random marker selection. The curve corresponding to random marker selection is shown in grey and is the lowest (worst) curve in all four plots. The ARI score is shown in (a), the AMI score is shown in (b), and the Fowlkes-Mallows score is shown in (c). (d) contains the FM scores for larger numbers of markers selected, showing the scVI does approach the behavior of random marker selection in this case. Each clustering is carried out using 5-fold cross validation and scores are averaged across folds.

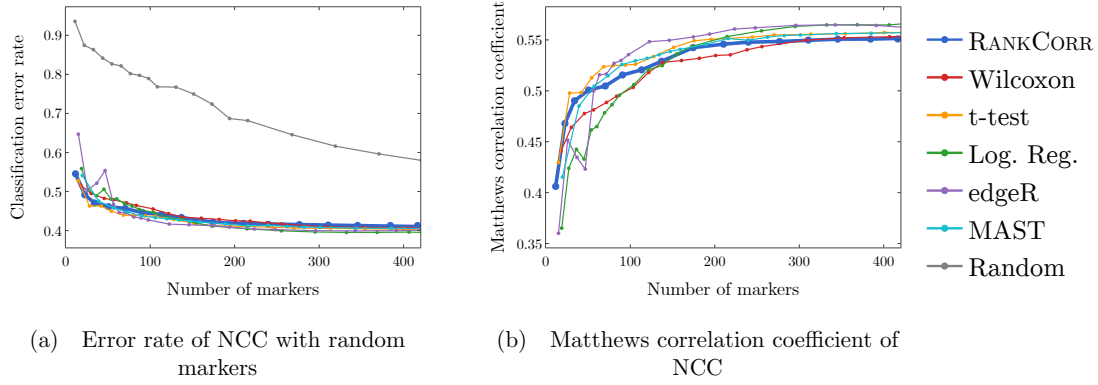


Figure 8: Supervised classification metrics for the ZHENGFI LT data set using the nearest centroid classifier (NCC). (a) is the classification error rate including a curve corresponding to random marker selection (the curve that starts at around 90% error). (b) contains the Matthews correlation coefficient. Data from random marker selection are not included in (b). Note that (b) is similar in shape to the classification accuracy ($1 - \text{classification error rate}$ from Figure 8(a) of the main manuscript).

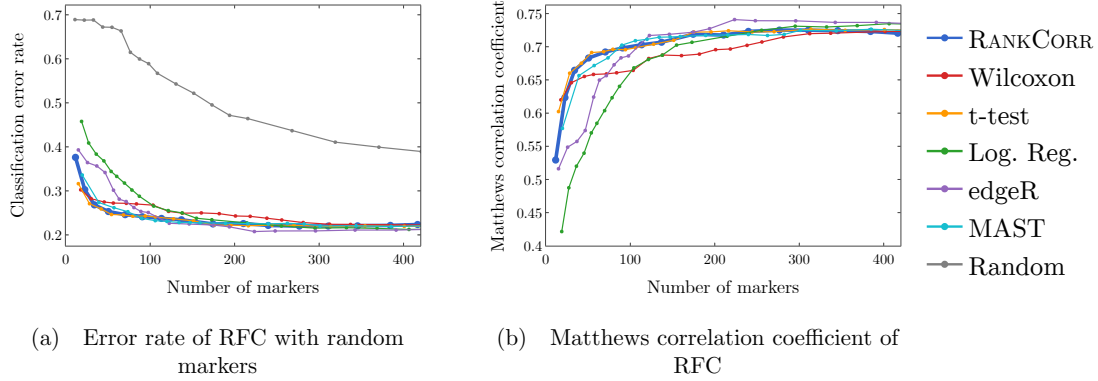


Figure 9: Supervised classification metrics for the ZHENGFI LT data set using the random forests classifier (RFC). (a) is the classification error rate including a curve corresponding to random marker selection (the curve that starts at around 70% error). (b) contains the Matthews correlation coefficient. Data from random marker selection are not included in (b). Note that (b) is similar in shape to the classification accuracy ($1 - \text{classification error rate}$ from Figure 9(a) of the main manuscript).

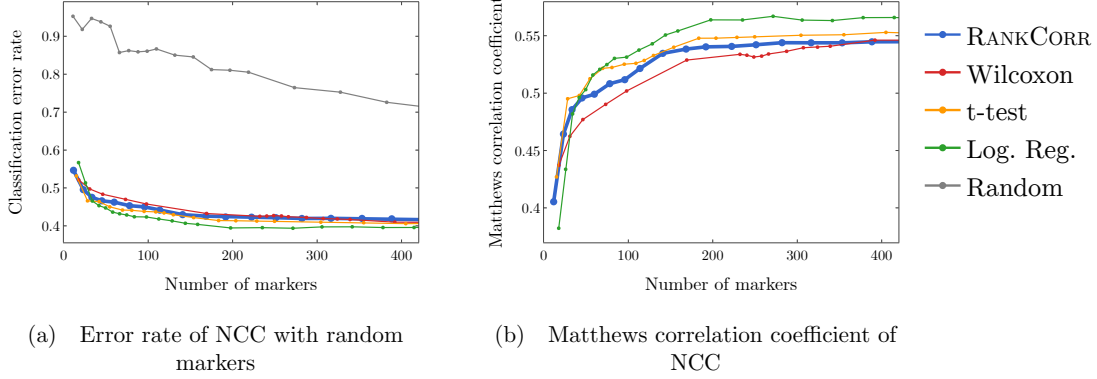


Figure 10: Supervised classification metrics for the ZHENGFULL data set using the nearest centroid classifier (NCC). (a) is the classification error rate including a curve corresponding to random marker selection (the curve that starts at around 90% error). (b) contains the Matthews correlation coefficient. Data from random marker selection are not included in (b). Note that (b) is similar in shape to the classification accuracy ($1 - \text{classification error rate}$ from Figure 8(c) of the main manuscript).

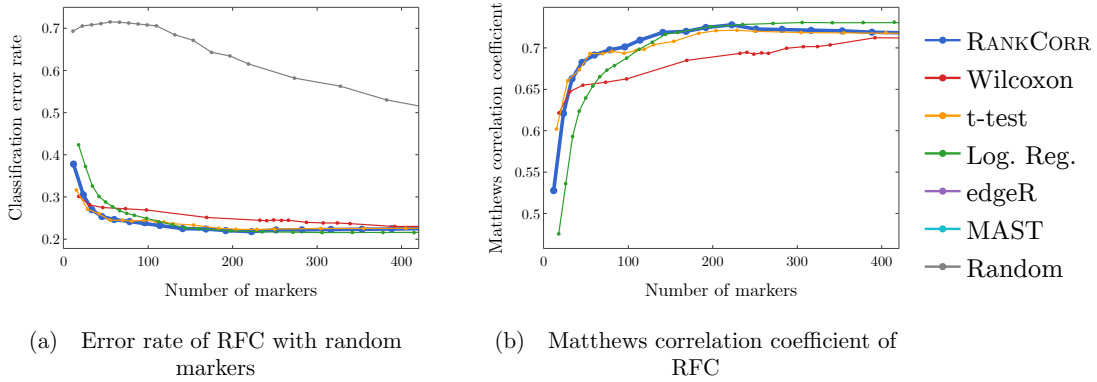


Figure 11: Supervised classification metrics for the ZHENGFULL data set using the random forests classifier (RFC). (a) is the classification error rate including a curve corresponding to random marker selection (the curve that starts at around 70% error). (b) contains the Matthews correlation coefficient. Data from random marker selection are not included in (b). Note that (b) is similar in shape to the classification accuracy ($1 - \text{classification error rate}$ from Figure 9(c) of the main manuscript).

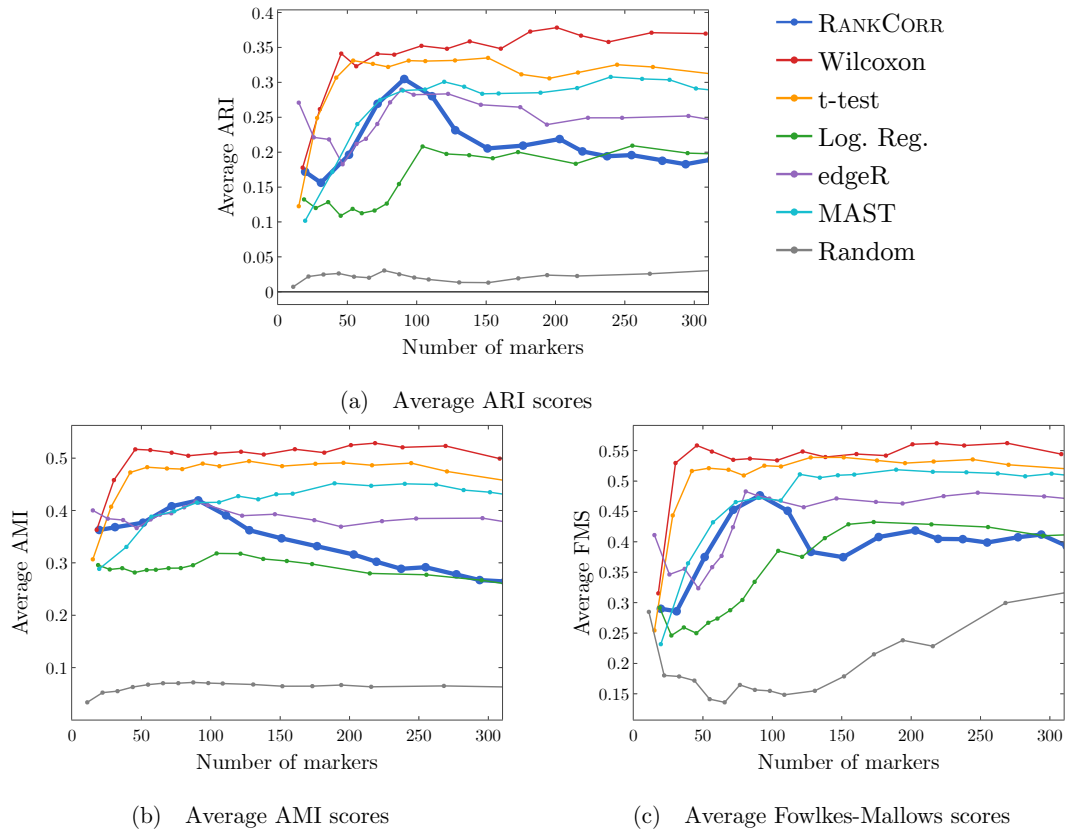


Figure 12: Unsupervised clustering metrics for the ZHENGFilter data set including data from random marker selection. The curve corresponding to random marker selection is shown in grey and is the lowest (worst) curve in all three plots. The ARI score is shown in (a), the AMI score is shown in (b), and the Fowlkes-Mallows score is shown in (c). Each clustering is carried out using 5-fold cross validation and scores are averaged across folds.

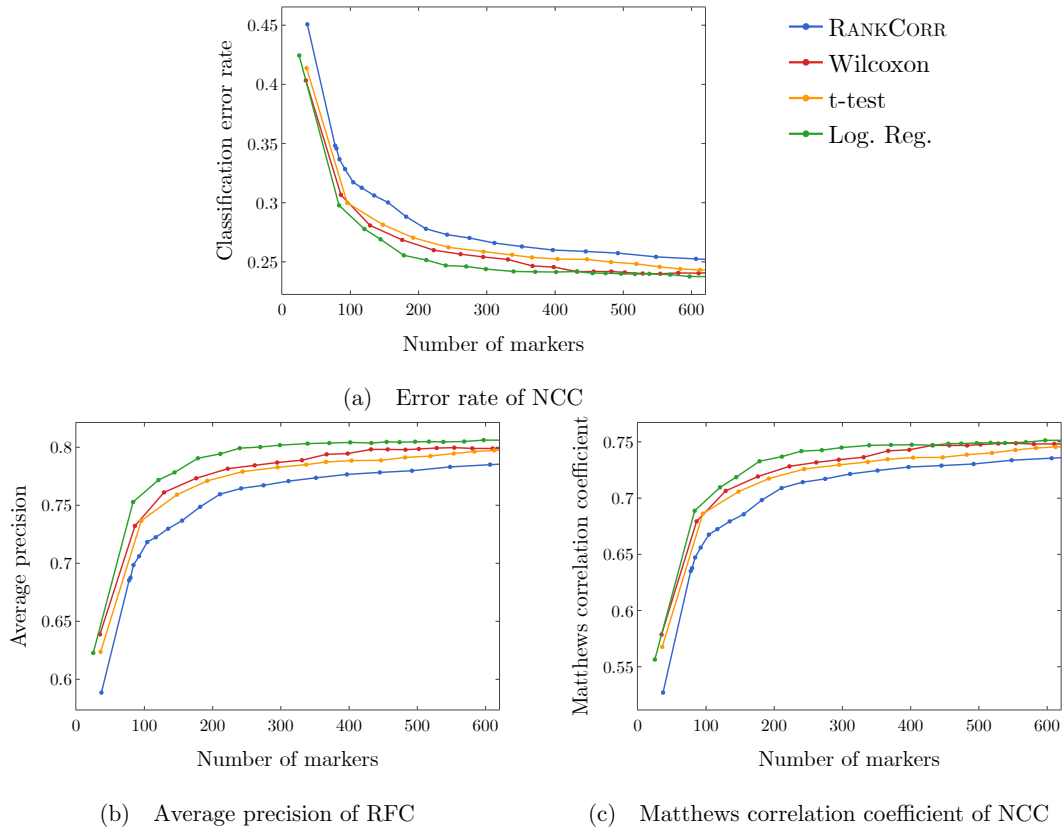


Figure 13: Supervised classification metrics for the 10xMOUSE data set using the nearest centroid classifier (NCC). (a) contains the classification error rate curves, (b) contains the average precision curves, and (c) contains the Matthews correlation coefficient curves. We do not compare to random markers on this data set.

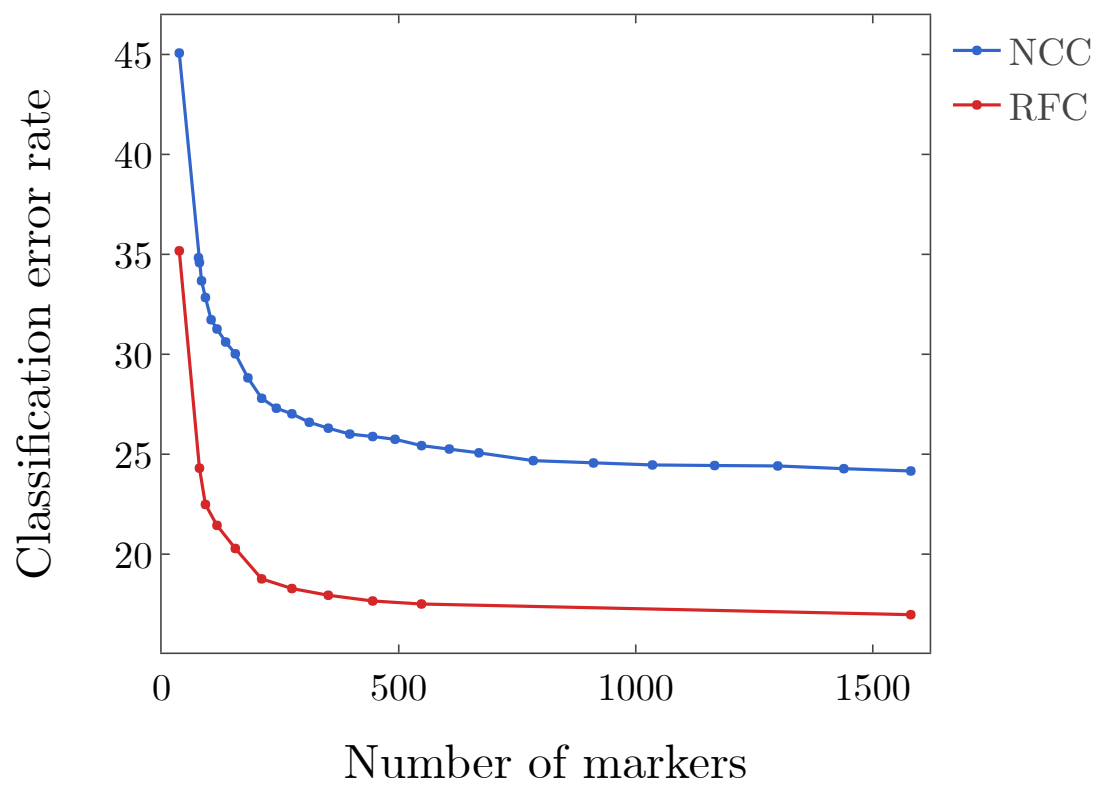


Figure 14: A comparison of the nearest centroid classifier (NCC) and the random forest classifier (RFC) using the RANKCORR method on the 10xMOUSE data set

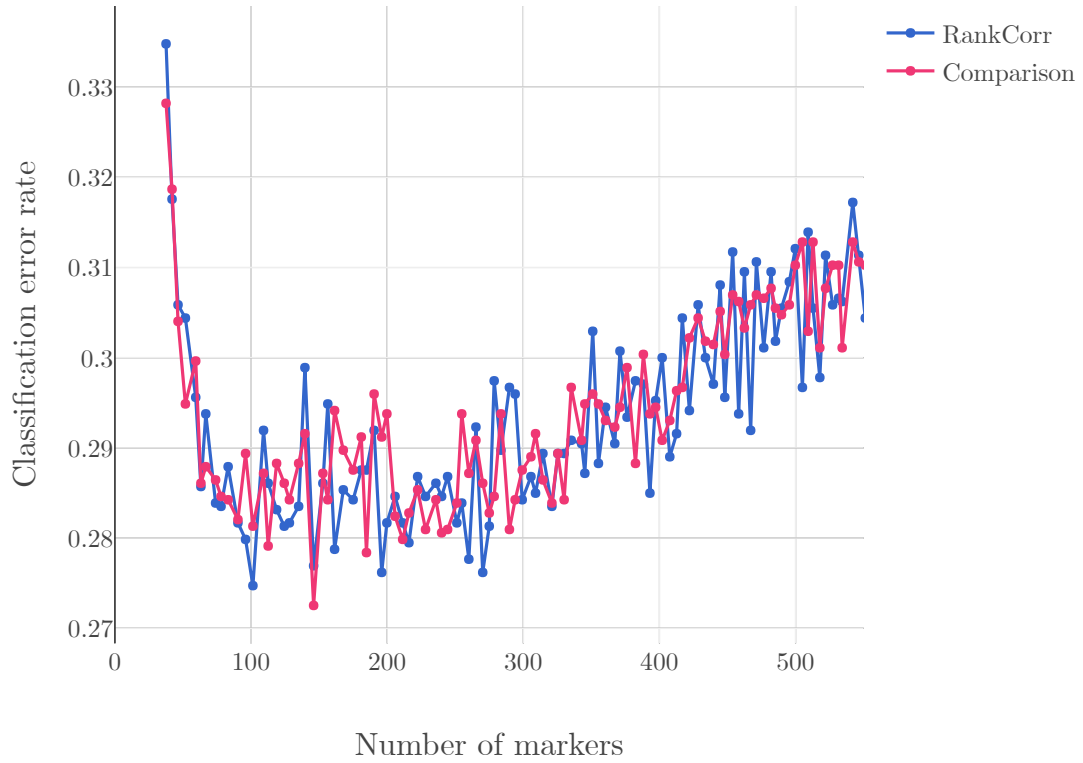


Figure 15: The classification accuracy under the RFC on the PAUL data set (see the Methods in the main manuscript) run twice with the same markers used for each point. Significant variation is observed in the classification accuracy over the two classification attempts. Differences of nearly 2% are observed between the two curves.

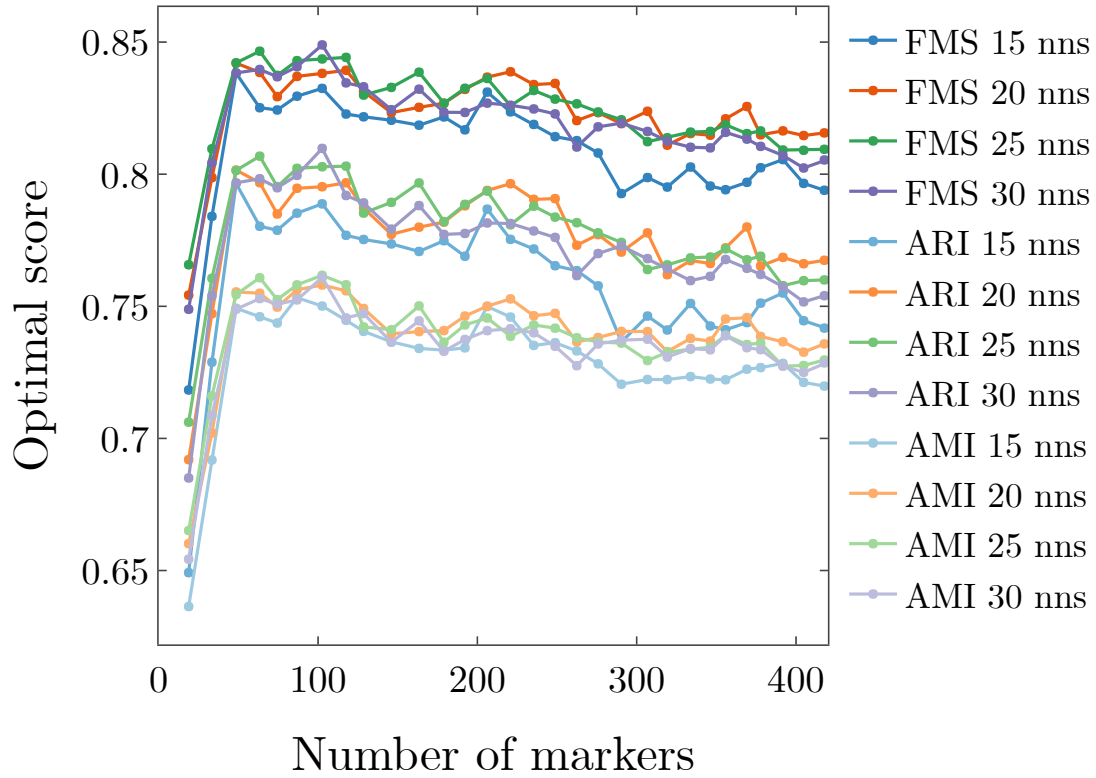


Figure 16: Effect of changing the number of nearest neighbors on the ARI, AMI, and FM scores for the ZEISEL data set using RANKCORR to select markers. Clustering was performed with Louvain and the scores were optimized over the resolution. It appears that 15 nearest neighbors is too few, while 30 nearest neighbors is too many.

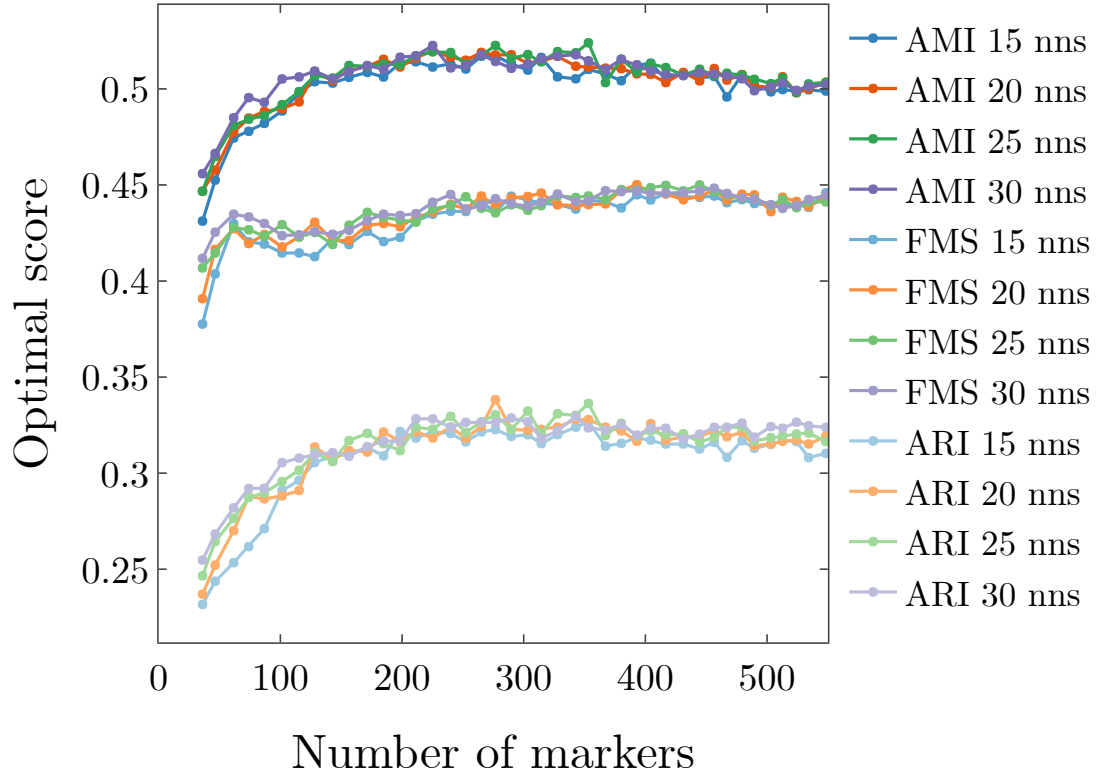
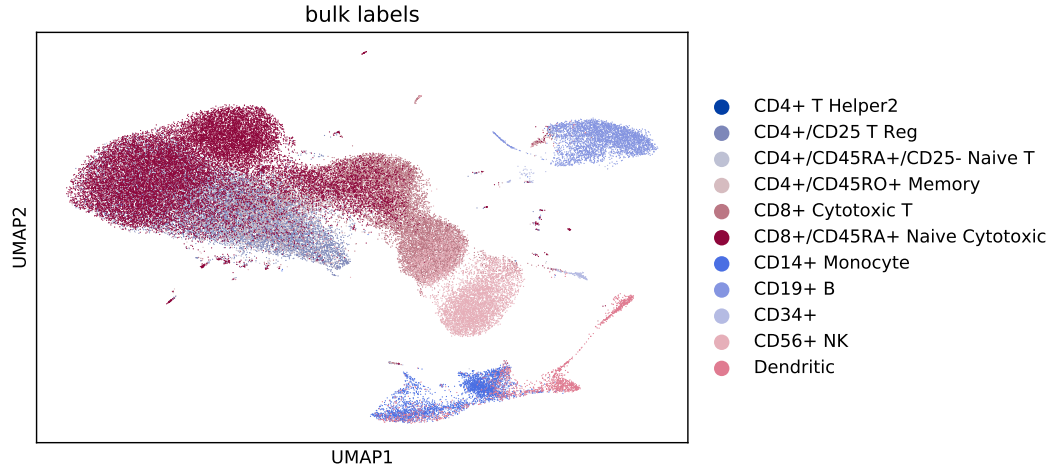
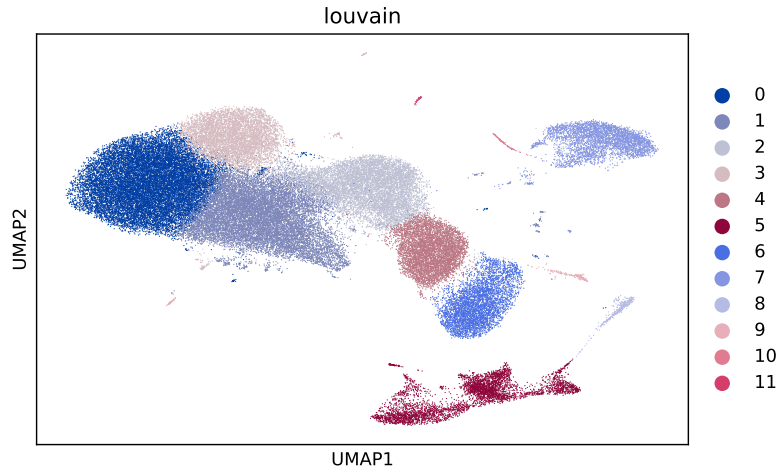


Figure 17: Effect of changing the number of nearest neighbors on the ARI, AMI, and FM scores for the PAUL data set using RANKCORR to select markers. Clustering was performed with Louvain and the scores were optimized over the resolution. All of the choices of numbers of nearest neighbors produce similar curves for all three scores. Choosing 30 nearest neighbors appears to provide increased performance for small numbers of markers.



(a)



(b)

Figure 18: Clustering the 68k PBMC data set from [1] (reference [2] from the main manuscript) with Louvain clustering. (a) contains a UMAP plot of the bulk labels. (b) is a UMAP plot of a Louvain clustering of the data set. It was created by first filtering to the 1000 most variable genes using the `cell_ranger` flavor of the `filter_genes_dispersion` function in the `scanpy` python package. The Louvain algorithm was run on the top 50 PCs and used 25 nearest neighbours for each cell with a resolution parameter of 0.3. The Louvain clustering solution subjectively looks similar to the bulk labels. The ARI for the clustering compared to the bulk labels is 0.345, the AMI is 0.565, and the FMS is 0.462 (these values have been rounded to 3 significant digits).

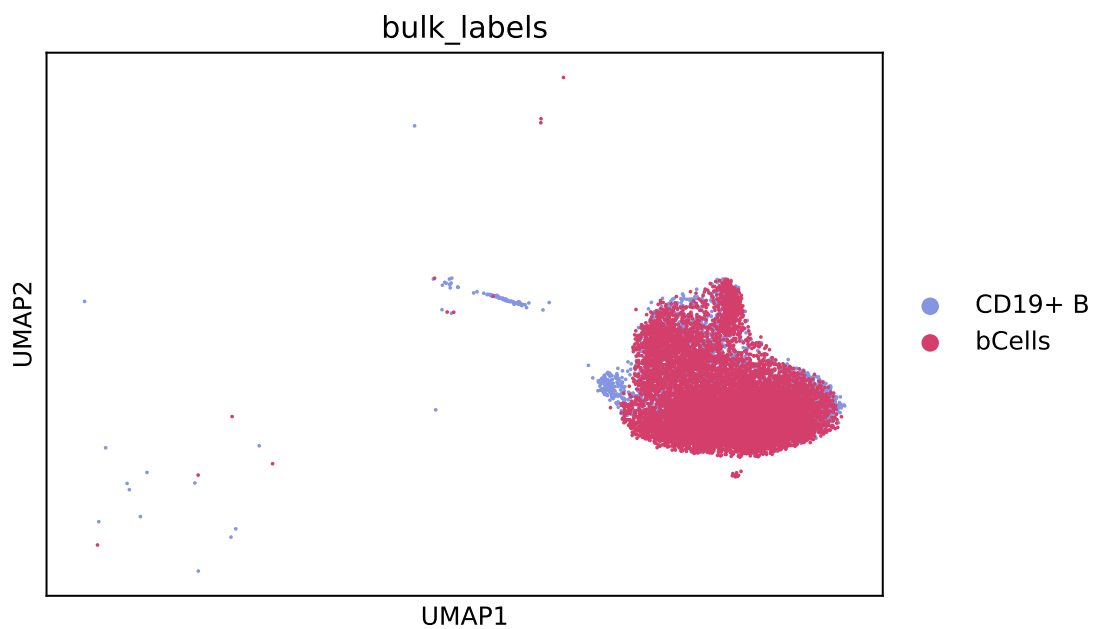


Figure 19: UMAP projection of the data consisting of ZHENGFULL combined with the isolated CD19+ B cell data set from [1] (reference [2] from the main manuscript) that was used to estimate parameters in Splatter simulations for generating synthetic data. We show only the isolated CD19+ sample (labeled “bCells”) and the cluster of B cells from ZHENGFULL. The overlap between the two clusters is quite good.

References

- [1] Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., Gregory, M.T., Shuga, J., Montesclaros, L., Underwood, J.G., Masquelier, D.A., Nishimura, S.Y., Schnall-Levin, M., Wyatt, P.W., Hindson, C.M., Bharadwaj, R., Wong, A., Ness, K.D., Beppu, L.W., Deeg, H.J., McFarland, C., Loeb, K.R., Valente, W.J., Ericson, N.G., Stevens, E.A., Radich, J.P., Mikkelsen, T.S., Hindson, B.J., Bielas, J.H.: Massively parallel digital transcriptional profiling of single cells. *Nature Communications* **8**, 14049 (2017). Article