SOCIETY FOR NATURAL SCIENCES STUDENT JOURNAL

Society for Natural Sciences Student Conference
University of Bath 8-9 April 2026

Assessing Statistical Methods for Clustering and Differential Gene Expression in scRNA-seq Data

Tabitha Percy

Single-cell RNA sequencing (scRNA-seq) is a field of transcriptomics that can provide information about cell heterogeneity, defined as the differences between cell states or types caused by changes during cell growth[1]. Differential gene expression (DGE) is crucial in evaluating the genomic changes that drive heterogeneity. As a result, improving confidence in estimates of highly expressed genes becomes of high importance in scRNA-seq data analysis.

A common pitfall in determining these genes is the inflation of the type I error rate, where p-values are artificially low due to 'double-dipping', occurring when the same data are used for clustering and DGE. In the clustering stage, two methods - Countsplit and SVMs - are tested using a simulated scRNA-seq dataset with known parameters. Two DGE tests - MAST and Nebula - offer an alternative post-clustering approach to controlling type I error inflation.

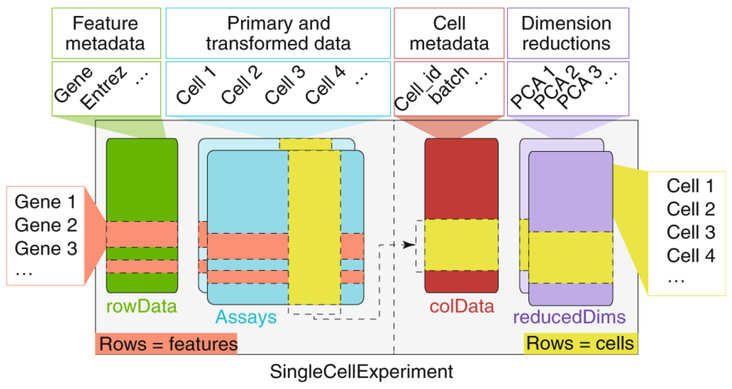


Figure 1: Schematic showing the data contained within a SingleCellExperiment (SCE) object in R. Figure taken from [2].

The SingleCellExperiment object consists of a cell by gene matrix with multiple features contained within the cell and gene (feature) metadata as seen in Figure 1. These features may be of interest during analysis, as they may interact with clustering and DGE results.

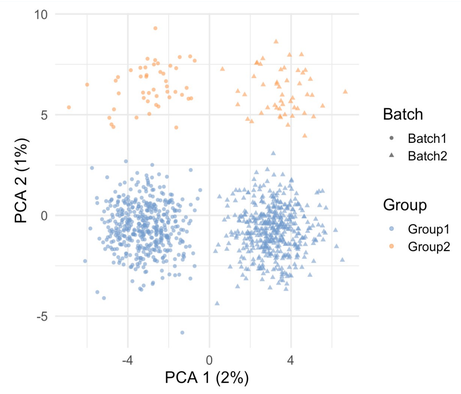


Figure 2: A PCA scatter plot produced in R with simulated parameters of 1:9 cells in each group and equal numbers of cells in each batch via Splatter.

Splatter is an R library that generates SCE objects with specified parameters[3]. Figure 2 shows how batches, in cell metadata, can interfere with clustering in a PCA plot.

Three methods are proposed for the clustering stage:

Naïve $X_{train} = X_{test}$

The same dataset is used for clustering (train) and DGE (test).

Countsplit $X_{train} \sim NB(X, 훍 = X - X_{test}$

Training and test datasets are split by counts as a form of k-fold cross-validation[4].

Cell Subsetting with SVMs

Training and test datasets are split by rows using support vector machines to guide feature selection[5].

A clustering algorithm, i.e. KNN, Leiden, Hierarchical, is then applied to the dataset to estimate clusters after conducting a principal components analysis.

These three methods were tested on a Splatter-simulated SCE with group effects. The adjusted Rand index (ARI) is returned to measure the similarity of clusters governed by group effects and estimated clusters.

Given that the algorithmic clustering of the dataset is a random process, code in R to calculate ARI was ran 50 times to calculate a 95% confidence interval for each method:

Method	Mean	95% C.I.
Naive	0.946	[0.918, 0.974]
Countsplit	0.724	[0.665, 0.782]
Cell Subsetting	0.922	[0.868, 0.976]

Table 1: Mean values and 95% confidence intervals for the Adjusted Rand Index returned from 50 iterations of each clustering method on the same Splatter-simulated dataset. Leiden clustering[6] is used.

Table 1 demonstrates a lower ARI returned by countsplit, possibly as only half of the available cell counts are used in the training matrix when clustering. Additionally, less genes are used during Variance-stabilizing transformation[7] due to more outliers observed using countsplit.

Naïve clustering shows the highest ARI, followed closely by cell subsetting, which may be a result of clustering on a population that is not fully representative of the dataset.

An alternative method to reduce type I error inflation is in the post-clustering stage: Two DGE tests that resist this are as follows:

MAST (Model-based Analysis of Single-cell Transcriptomics)[8] fits generalised linear models to the dataset, adapted for bimodal or low-count data, which often lead to higher type I error rate.

Nebula (NEgative Binomial mixed model Using a Large-sample Approximation)[9] provides a faster method of inference about the features in a cell population and relies on large sample sizes.

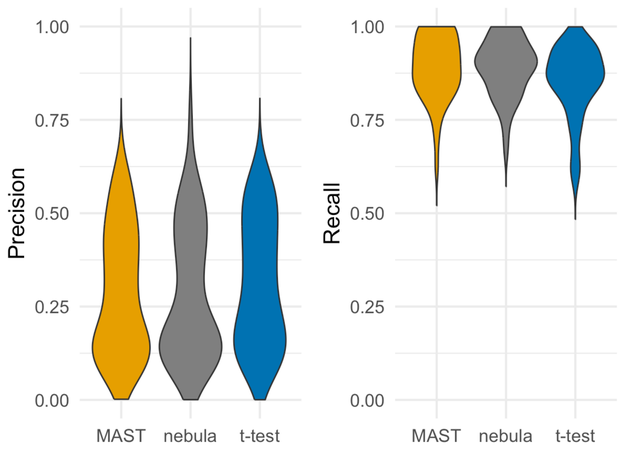


Figure 3(a): A violin plot of precision and recall metrics returned by naïve Leiden clustering, coloured by choice of DGE test. Values obtained by running DGE tests 50 times.

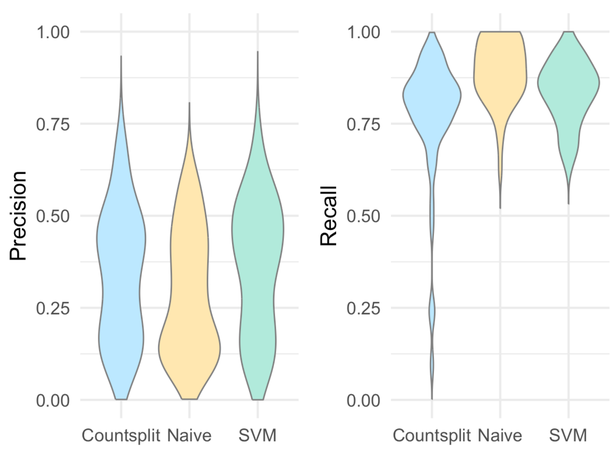


Figure 3(b): A violin plot of precision and recall metrics returned by naïve, countsplit, and cell subsetting (SVM) paired with Leiden clustering, coloured by choice of method. Values obtained by running clustering methods with nebula 50 times.

Figure 3(a) depicts higher precision and recall shown from MAST and nebula compared to a t-test in naïve clustering, indicating that these methods are more suitable to simulate clusters in a Splatter SCE. Nebula shows less variation in the mean values of both precision and recall than MAST, which provides evidence that it is consistently estimating simulated clusters.
Figure 3(b) shows the large range in precision and recall values returned by countsplit, which may be due to similar reasons for its low ARI. Cell subsetting with nebula returns the highest precision and recall values.

The bimodal distribution seen in figures 3(a) and 3(b) is likely a result of overclustering of certain iterations, which leads to lower precision and recall. Leiden clustering is conducted on a K-nearest neighbours graph from the dataset, which prevents fixing the number of estimated clusters to match the number of simulated groups in the Splatter SCE. Another clustering algorithm, such as k-means or hierarchical, must be used in this case. However, these algorithms may make assumptions about the data that drive inflation of type I error.

Rank	Method combination	Mean Precision	Mean Recall
1	MAST (Cell subsetting)	0.44 +/- 0.05	0.80 +/- 0.03
2	nebula (Countsplit 1:5 split)	0.39 +/- 0.04	0.73 +/- 0.05
3	nebula (Countsplit 1:1 split, small group)	0.38 +/- 0.04	0.73 +/- 0.04
4	nebula (Cell subsetting)	0.36 +/- 0.05	0.84 +/- 0.02
5	nebula (Countsplit 1:1 split)	0.34 +/- 0.05	0.77 +/- 0.05
6	nebula (Cell subsetting, small group)	0.33 +/- 0.04	0.83 +/- 0.03
7	MAST (naive clustering)	0.31 +/- 0.05	0.85 +/- 0.03
8	nebula (naive clustering)	0.29 +/- 0.05	0.89 +/- 0.02
9	t-test (naive clustering)	0.29 +/- 0.05	0.89 +/- 0.02
10	nebula (Cell subsetting, low cell counts)	0.16 +/- 0.05	0.06 +/- 0.02
11	nebula (Countsplit 1:1 split, low cell counts)	0.00 +/- 0.00	0.00 +/- 0.00

Table 2: Complete set of tested method combinations with mean precision and recall values, with uncertainty dictated by half the range of the data. Summary statistics were obtained by running each method combination 50 times. Ranking is dictated by the mean precision value returned.

Table 2 reinforces the conclusions from figures 3(a) and 3(b), providing background with other tested combinations. Extensive testing concludes that the most suitable methods to estimate clusters are cell subsetting used with MAST, and a 1:5 Countsplit on train:test count matrices paired with nebula. Given that nebula assumes larger sample sizes, and MAST is suited to bimodal data, these conclusions are in line with what we expect.

[1] Jovic, D., Liang, X., Zeng, H., Lin, L., Xu, F., & Luo, Y. Single-cell RNA sequencing technologies and applications: A brief overview. *Clin. Transl. Med.*, **3**, 12 (2022).
[2] Amezquita, R.A., Lun, A.T.L., Becht, E. et al. Orchestrating single-cell analysis with Bioconductor. *Nat. Methods* 17, 137–145 (2020).
[3] L. Zappia, B. Phipson, A. Oschlack. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.* **18**, 174 (2017).
[4] A. Neufeld, L.L .Gao, J. Popp, A. Battle, D. Witten. Inference after latent variable estimation for single-cell RNA sequencing data, *Biostatistics*, **25**, 270 (2024).
[5] W.S. Noble. What is a support vector machine? *Nat Biotechnol.* **24**, 1565 (2006).
[6] Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep.* **9,** 5233 (2019).
[7] Bartlett, M. S. The Use of Transformations. *Biometrics* **3,** 39 (1947).
[8] Finak, G., McDavid, A., Yajima, M. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).
[9] He, L., Davila-Velderrain, J., Sumida, T.S. *et al.* NEBULA is a fast negative binomial mixed model for differential or co-expression analysis of large-scale multi-subject single-cell data. *Commun. Biol.* **4**, 629 (2021).

Special thanks to:
Dr. Magdalena Strauss at The University of Exeter Living Systems Institute.

Funded by the #ExeterForever ALUMNI ANNUAL FUND