Denoising scRNA-seq annotations using deep learning

Low Boon How
Nanyang Technological University
blow011@e.ntu.edu.sg

Hosted by Nanjing University
Supervised by Prof. Chen Dijun

Luke Price
University of Leeds
sc22lip@leeds.ac.uk

Abstract

Conventional scRNA-seq analysis methods can often be labor-intensive and prone to human biases. Here, we will investigate a different approach, where we train a Variational Autoencoder (VAE) to aid with cell annotation via data reconstruction. The VAE seeks to denoise scRNA-seq data to produce better results when using existing automated annotation methods.

Context: Conventional scRNA-seq Methods

Basic scRNA-seq analysis requires multiple steps of data processing and human annotations. These steps are often tedious and require domain expertise. **(Fig. 1)**

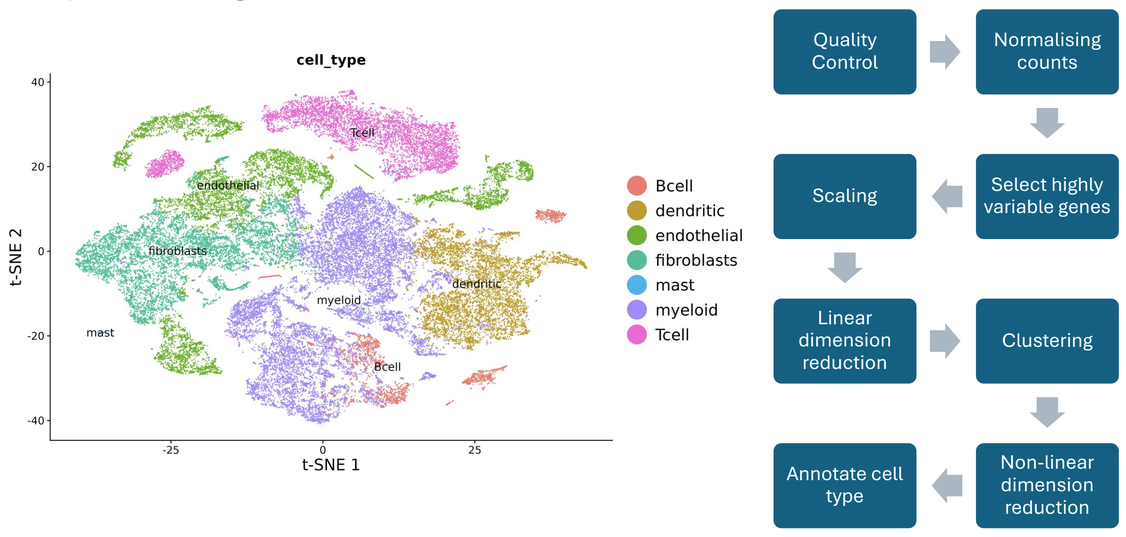

Fig. 1: A visualization of the standard scRNA-seq workflow with completed annotations

Many methods exist to ease the process of scRNA-seq analysis, one of which is **single-cell Variational Inference (scVI).**

Literature review: scVI

scVI is a conditional Variational Autoencoder (cVAE) designed to aid with tasks including **batch correction** or **visualization. (Fig. 2)** It models each gene in a cell as a sample drawn from a **zero-inflated negative binomial distribution (ZINB).**

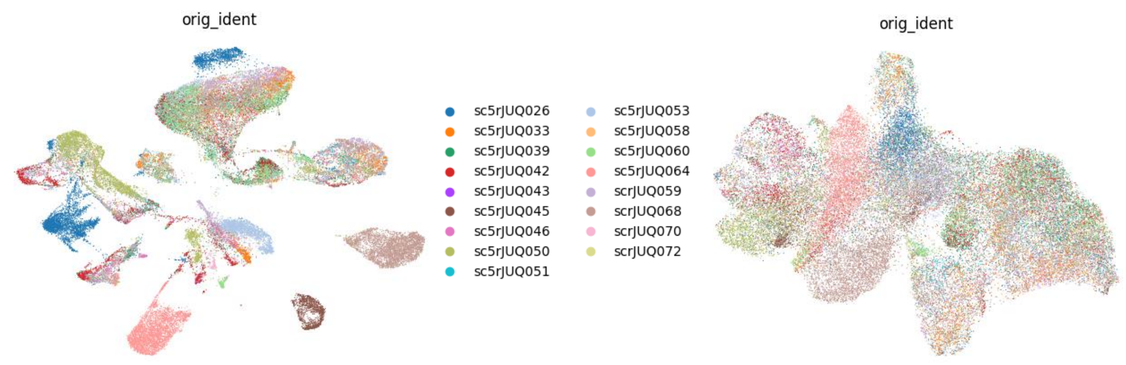

Fig. 2: UMAP of scRNA-seq counts before and after integration was done using scVI

ZINB accounts for the inflated zeros in RNA-seq by having **different Probability Density Functions (PDF) for zero and non-zero counts**. The overall PDF is given as such, where θ and μ are the mean and dispersion parameters, respectively. **(Eqn. 1)**

$$P(X = x) = \begin{cases} \pi + (1 - \pi)\, g(x = 0) & \text{if } x = 0, \\ (1 - \pi)\, g(x) & \text{if } x > 0 \end{cases}$$

$$\text{where } g(x, \mu, \theta) = \frac{\Gamma(x+\theta)}{x!\Gamma(\theta)} \left(\frac{\theta}{\theta+\mu}\right)^{\theta} \left(\frac{\mu}{\theta+\mu}\right)^{\mu}$$

Eqn. 1: Probability Distribution Function of a Zero-inflated Negative Binomial distribution

Methods: Modelling

Conditional VAEs (cVAE) like scVI **require additional neurons to train their conditions (Fig. 3)**, making them **more computationally expensive**. Thus, we will look at a simpler model, a **vanilla VAE (Fig. 4)**, to conduct our study. Our VAE will also utilize a **modified sampling method inspired by ZINB**. **(Fig. 5, 6 and Eqn. 2, 3)**

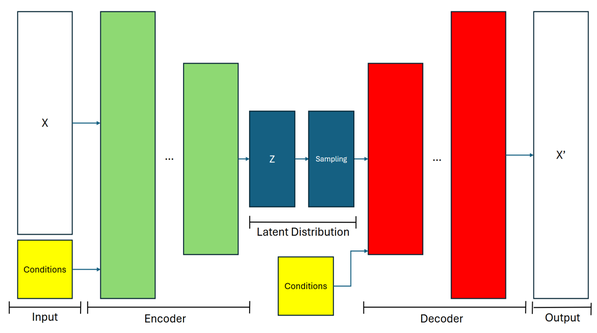

Fig. 3: Architecture of cVAE

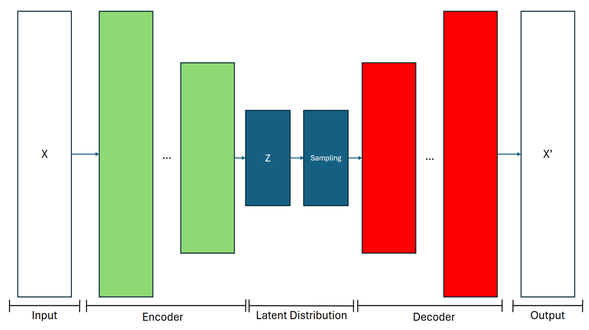

Fig. 4: Architecture of vanilla VAE

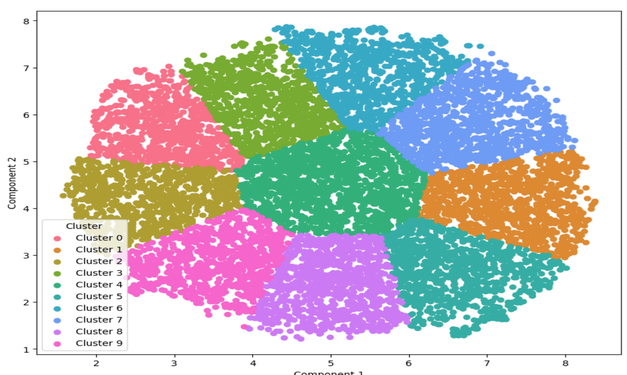

Fig. 5: UMAP of latent space using Eqn. 2

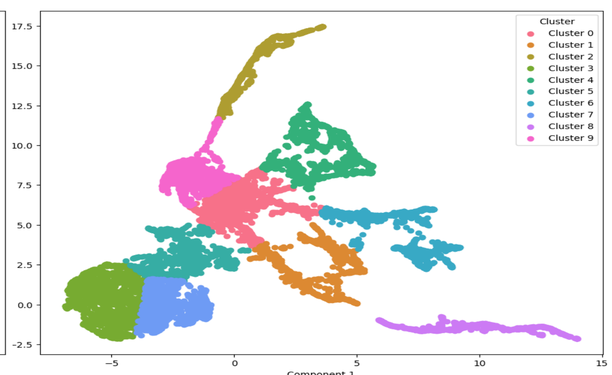

Fig. 6: UMAP of latent space using Eqn. 3

$$z = \mu + \sigma * \epsilon$$

μ is the mean
σ is the variance
ϵ is a sample of the distribution

Eqn. 2: Gaussian sampling

$$z = \mu + \sigma * \epsilon * (1 - \pi)$$

μ is the mean
σ is the variance
ϵ is a sample of the distribution
π is probability of zeros

Eqn. 3: Custom sampling

Methods: Data preprocessing and labelling

Highly Variable Genes (HVGs) selection is important to capture biologically significant features. We tested two methods of HVG selection:

1. default function using Seurat package in R **choosing 2000 genes.**
2. custom function where genes are only retained if they have a **log-normalized variance > 0.5** and **0.125 < log-normalized mean < 3**.

We labelled cell-types fully via the **singleR package in R.** We took the data from **EMBL-EBI Database accession E-MATB-8107.**

Results and discussions

We compared the SingleR labels across the same embedding space. **The number of cell types annotated had already been reduced.** Also, there is a significant reduction of noise within the clusters. **(Fig. 7)**

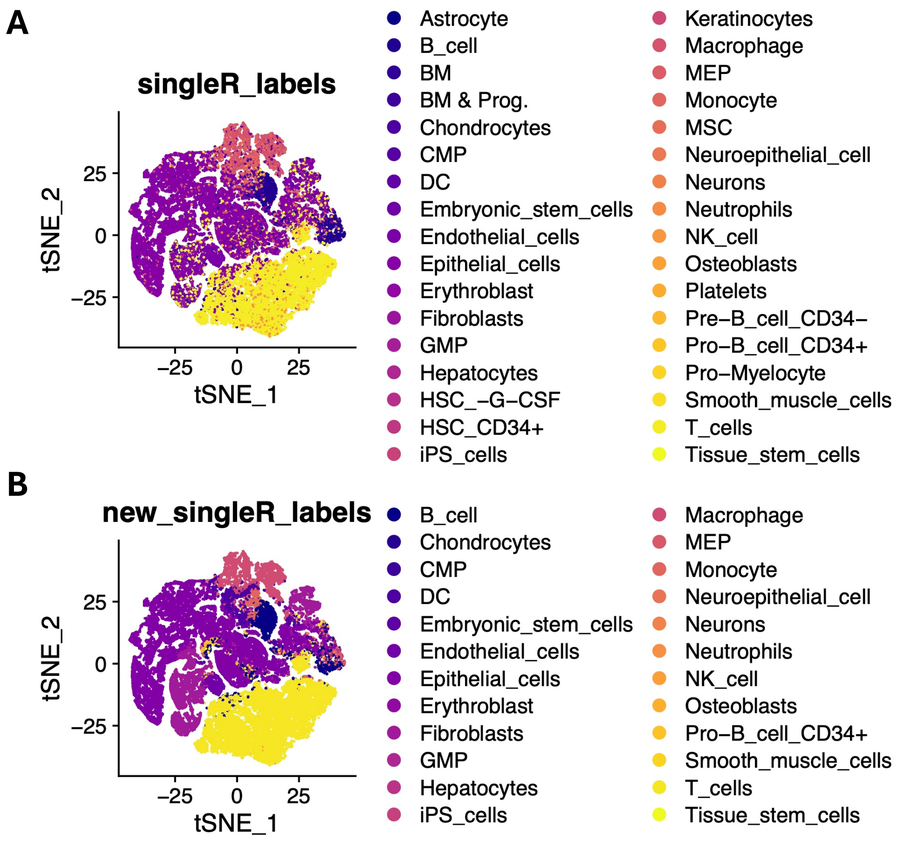

Fig. 7: Comparison of singleR labels before (A) and after (B) data reconstruction on the same t-SNE embedding

We plotted heatmaps of **cell type annotations per cluster** for all permutations of data preprocessing. **(Fig. 8)** We also tested hyperparameter tuning by **increasing the neurons at each model layer**. If a **column (cluster)** has **multiple rows (cell types)**, it indicates that the cluster is noisy as many **cell types are allocated**, and vice versa.

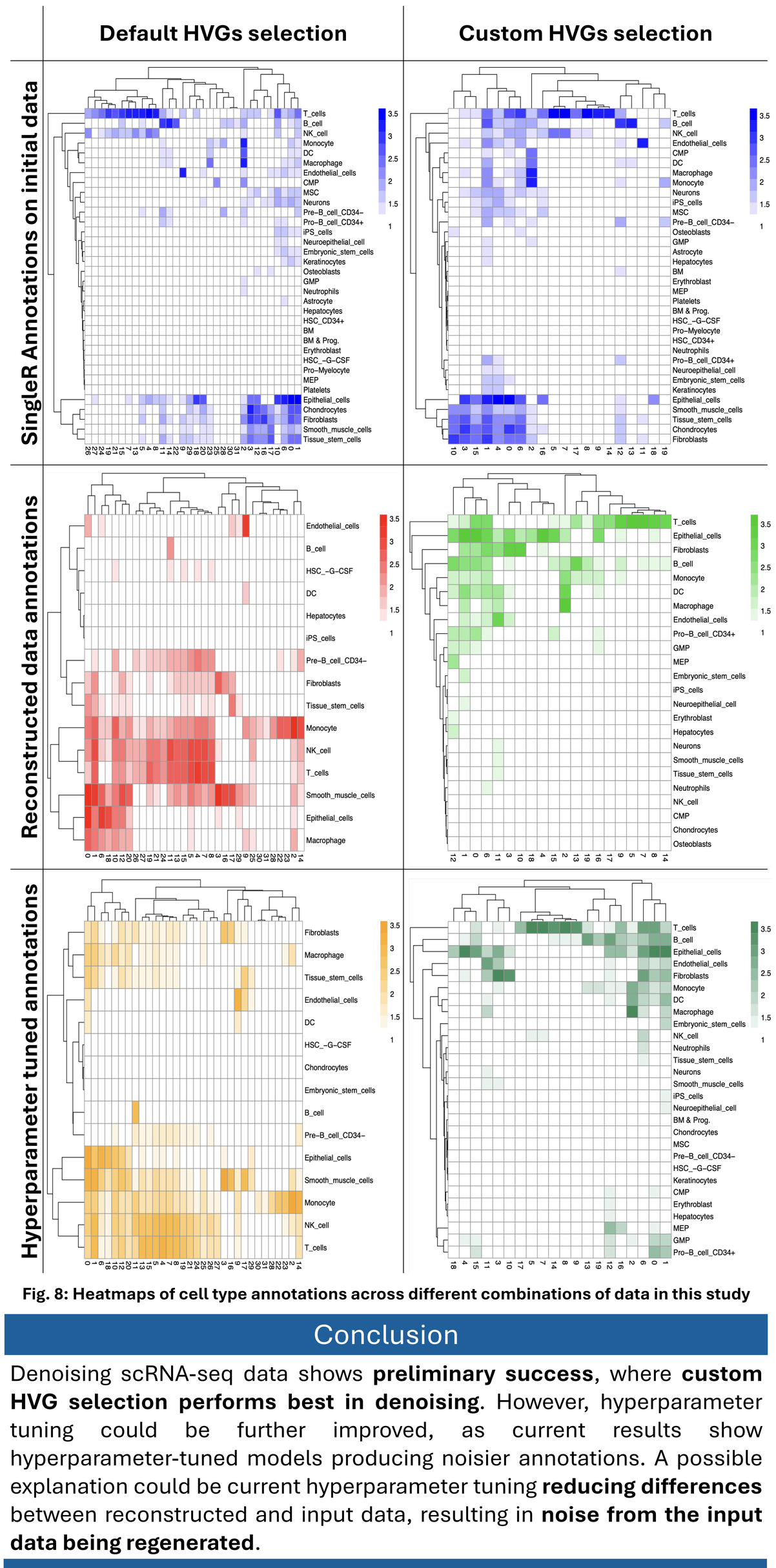

Fig. 8: Heatmaps of cell type annotations across different combinations of data in this study

Conclusion

Denoising scRNA-seq data shows **preliminary success**, where **custom HVG selection performs best in denoising**. However, hyperparameter tuning could be further improved, as current results show hyperparameter-tuned models producing noisier annotations. A possible explanation could be current hyperparameter tuning **reducing differences** between reconstructed and input data, resulting in **noise from the input data being regenerated**.

References

Lopez, Romain, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. 2018. "Deep Generative Modeling for Single-Cell Transcriptomics." *Nature Methods* 15 (12): 1053–58. https://doi.org/10.1038/s41592-018-0229-2.

Jiang, Ruochen, Tianyi Sun, Dongyuan Song, and Jingyi Jessica Li. 2022. "Statistics or Biology: The Zero-Inflation Controversy about ScRNA-Seq Data." *Genome Biology* 23 (1). https://doi.org/10.1186/s13059-022-02601-5.

Aran, Dvir, Agnieszka P. Looney, Leqian Liu, Esther Wu, Valerie Fong, Austin Hsu, Suzanna Chak, et al. 2019. "Reference-Based Analysis of Lung Single-Cell Sequencing Reveals a Transitional Profibrotic Macrophage." *Nature Immunology* 20 (2): 163–72. https://doi.org/10.1038/s41590-018-0276-y.

Qian, Junbin, Siel Olbrecht, Bram Boeckx, Hanne Vos, Damya Laoui, Emre Etlioglu, Els Wauters, et al. 2020. "A Pan-Cancer Blueprint of the Heterogeneous Tumor Microenvironment Revealed by Single-Cell Profiling." *Cell Research* 30 (9): 745–62. https://doi.org/10.1038/s41422-020-0355-0.

Carlo De Donno et al., "Population-Level Integration of Single-Cell Datasets Enables Multi-Scale Analysis across Samples," *Nature Methods* 20, no. 11 (October 9, 2023): 1683–92, https://doi.org/10.1038/s41592-023-02035-2.