Deep generative approach to model single-cell expression data of human embryoid bodies

G Prashant (BS17B011), Shreya Nema (BS17B033)

BT5240 - Computational Systems Biology

Abstract: The advancements in next generation sequencing technologies have resulted in massive amounts of genomic and transcriptomic data being generated. With the emergence of single-cell RNA-sequencing technologies, it is now possible to track the complex heterogeneous behaviour within cell populations, allowing us to get a single-cell picture of a variety of phenotypes. Several studies have analysed and profiled the gene expression levels of cell types of pluripotent human embryonic stem cells (hESCs). If hESCs clump together, they form embryoid bodies, which have the ability to differentiate spontaneously into muscle cells, nerve cells, and other cell types. In this project, we intend to explore deep generative modelling of single-cell data of human embryoid bodies using variational autoencoders. The encoding layers reduce the dimensions of the data to generate a latent representation of a cell that captures the most essential features. The decoding layers reconstruct the data from the latent representation. Collectively, we perform clustering, visualisation and differential expression to characterise subpopulations that can differentiate into specific specialised cell types based on predicted marker genes. We manually label six predicted cell clusters to major progenitor cell types, including stromal cells, muscle cells, endothelial cells, neural cells, epithelial cells and liver cells using differentially expressed genes. In the process, we also highlight marker genes that are enriched in each cluster.

Keywords: hESC, scRNA sequencing, embryoid bodies, deep generative models, autoencoders

INTRODUCTION

Embryonic stem cells and embryoid bodies: A single-cell perspective

Stem cells are undeveloped and unspecialised cells that have the potential to divide and become new stem cells or differentiate into specialised cell types with specific functions. The primary role of adult stem cells that are found among differentiated cells in tissues is to repair and replenish dying cells, thereby maintaining homeostasis. On the other hand, embryonic stem cells are cells that are derived from the inner mass cells of an embryo. Human embryonic stem cells (hESCs) can be utilised to address fundamental questions in multiple fields of regenerative medicine, advanced cancer research, developmental biology and disease modelling. They are pluripotent and can differentiate into all the derivatives of the three primary germ layers - ectoderm, endoderm, and mesoderm, all of which can generate more than 200 specialised cell types in total. hESCs are also capable of indefinitely self-renewing in an undifferentiated state under appropriate conditions. While the nature and source of variability are poorly understood, it becomes essential to identify the molecular mechanisms that are responsible for controlling hESC pluripotency and self-renewal. This heterogeneity in cells has led to practical impediments in research. The gene expression program of hESCs allows them to self-renew and also assures their differentiation into several cell types in response to specific developmental cues. These cues include core regulatory circuitry that maintains hESCs and the genes that encode lineage-specific transcription factors. Cultured hESCs clump together and form three-dimensional aggregates known as embryoid bodies (EB). Formation of EB initiates spontaneous differentiation towards the three germ lineages.

With the advancements in high-throughput single-cell RNA sequencing technologies, it is possible to capture cellular heterogeneity at the level of a single cell. This heterogeneity can be further analysed to determine novel hESC-specific genes that could serve as reliable hESC markers associated with characteristic stem cell phenotype. Data analysis of gene expression patterns and functional classification can reveal characterised human EB differentiation status and can be used to determine the differences in various developmental events. This analysis of gene expression data includes preprocessing - quality control, normalisation, data correction, feature selection, and dimensionality reduction and cell- and gene-level downstream analysis. The inherent noise factors in scRNA-seq datasets include batch and dropout effects that must also be characterised to reflect the true nature of the data. Probabilistic modelling of cell populations allows multi-scale analysis of single-cell datasets across samples, subpopulations, and individual cells. Many experimental protocols and computational analysis approaches exist for single-cell RNA sequencing. Some of the popular methods

[⋆] Instructor: Dr Karthik Raman, IIT Madras include Seurat, Scanpy and Scater that provides large analysis toolboxes. With an increasing number of techniques to analyse single-cell data, it becomes important to navigate through these existing methods to better understand their potential drawbacks when handling large and real-world single-cell datasets.

Deep generative modelling

Over the past decade, deep learning techniques have proven to solve a multitude of complex problems. The increase in focus on leveraging deep learning tools to solve biological problems is attributed to the high dimensional and nonlinear nature of biological datasets. Autoencoders are artificial neural networks that encode data into a lower-dimensional latent representation, which is decoded to reconstruct the original data. They are typically used in machine translation, image denoising and segmentation. Deep generative modelling aims to learn the inherent distribution of the data using deep learning, unlike a discriminative approach that learns explicit boundaries between classes. Deep generative modelling can be implemented through Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs).

Variational Autoencoders

Several studies have implemented linear dimensional reduction techniques such as PCA on single-cell expression datasets in order to perform clustering and differential expression. However, as it is evident that expression data lies on a lower-dimensional nonlinear manifold in a high dimensional space, nonlinear dimensionality reduction techniques that capture the most essential features are more suitable for representing individual cells. Variational autoencoders achieve this by probabilistically encoding the expression levels of individual cells into a lower-dimensional latent space using artificial neural networks. The latent representation can further be decoded to reconstruct the original data with minimum loss. The latent representation can, therefore, be utilised for clustering and differential expression analysis.

A schematic diagram of a typical variational autoencoder is shown in Figure 1. The input data x is encoded to estimate the mean and covariance parameters of the Gaussian latent posterior distribution $q_\phi(z|x)$, from which the latent representation z is sampled, which is further decoded to reconstruct the input data $\hat{x}$. The reconstructed data is fit into a likelihood distribution $p_\theta(x|z)$. The modelling parameters ϕ, θ are estimated such that the negative data reconstruction loss is maximised and the KL divergence of the latent posterior $q_\phi(z|x)$ and latent prior $p(z)$ is minimised. The optimisation function, which is the variational lower bound of the marginalised log-likelihood of the data, is therefore given by

$$\mathcal{L} = E_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}[q_\phi(z|x)||p(z)]$$

The negative value of the marginalised log-likelihood is usually considered as the loss function to be minimised. It is referred to as Evidence Lower Bound (ELBO) loss and takes a positive value.

The latent prior $p(z)$ is generally assumed to be a standard Gaussian distribution with zero mean and unit covariance.

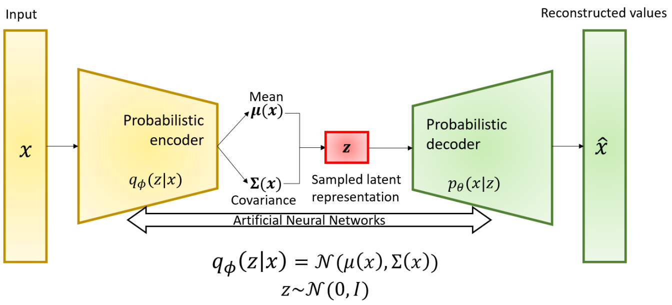

Fig. 1. Variational Autoencoder: A schematic; Reparameterization technique makes backpropagation possible by enabling the function to be differentiable at the bottleneck layer

The likelihood distribution $p_\theta(x|z)$ takes the form of a zero-inflated negative binomial (ZINB) distribution, which is popularly used to model gene counts. The negative binomial (NB) distribution is a discrete probability distribution for a random variable x which is given by

$$P(x) = NB(x) = \frac{\Gamma(x+\Theta)}{\Gamma(\Theta)} \left(\frac{\Theta}{\Theta+\mu}\right)^{\Theta} \left(\frac{\mu}{\Theta+\mu}\right)^{x}$$

The zero-inflated version of the negative binomial - ZINB is given by

$$P(x) = \pi\delta(x) + (1-\pi)NB(x)$$

$$\delta(x) = \begin{cases} 1 & x = 0 \\ 0 & x \neq 0 \end{cases}$$

where Θ and μ represent the gene dispersion and mean frequency respectively, and π represents the probability of a data entry being a dropout. Dropout effects occur due to the inability of the sequencing machine to capture the expression levels of lowly expressing mRNA transcripts. Dropouts result in an inflated number of zeros in the scRNA-seq data, thereby generating a highly sparse matrix with plenty of "false zeros".

In this project, we have implemented single-cell variational inference (scVI) package in python to perform the analyses on a single-cell RNAseq dataset of heterogenous day 8 human embryoid bodies. scVI takes into account and models the experimental noise such as batch effects and limited sensitivity that are inherent in scRNA expression data while preserving accurate estimates of variation in the data.

RESULTS

Model Selection and Clustering

Multiple models with varying hyperparameters were trained and the ELBO losses. The hyperparameters include latent size, number of hidden layers and number of hidden units. We trained 27 different combinations for 500 epochs by choosing latent sizes as 10, 20 and 30, symmetric number of hidden layers as 1, 2 and 3, and the number of hidden units in each layer as 128, 256 and 512. Upon analysing the ELBO losses of all the trained models, we noticed a significant decrease in the loss values as the number of hidden units in each layer increases for corresponding latent sizes and number of hidden layers, as illustrated in Figure 2.

Gaussian Mixture Models (GMMs) and k-Means clustering methods were implemented on the latent spaces of all the trained models. The number of cluster labels was chosen to be 6, in accordance with the previously done analyses on the dataset, that classifies these clusters into six major types of progenitor cells - muscle cells, stromal cells, endothelial cells, epithelial cells, liver cells and neural cells.

The performance of the clustering results of these models was characterised by a metric called silhouette score, which determines how well the clusters are separated from each other. Silhouette score is the average of the silhouette widths of all the data points. The silhouette width of a data point $i \in c_i$ is given by

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Here, $a(i)$ is the mean distance from i to all other data points in the same cluster c_i, and $b(i)$ is the least mean distance from i to all data points in all clusters but c_i.

It was observed that the silhouette scores of the clustering results performed with k-Means were better than those performed with GMM. This can be attributed to the probabilistic nature of Gaussian Mixture Models, unlike the distance-based approach of the k-Means algorithm. In addition to this, we also examined the relationship between the complexity of the model with the silhouette scores of their respective clustering results, as shown in Figure 3. We were able to notice that in some cases, despite having lower loss values, the models with higher complexity in terms of the latent size, number and size of hidden layers had lower values of silhouette scores when compared to less complex models. In other words,

Among the models that had an ELBO loss of less than 4200, the model that was trained with 512 hidden units, one hidden layer and a latent size of 10 had the highest silhouette score of 0.134 when clustered using k-Means. Hence, this model was chosen for differential expression analysis and cell type identification.

Visualisation

The most prevalent methods used to visualize higher dimensional data are PCA and t-Stochastic Neighborhood Embedding (t-SNE). However, in this project, we explore another visualisation technique called Uniform Manifold Approximation and Projection (UMAP). UMAP has recently proved to preserve the global structure of the data far better than t-SNE and PCA. It is widely used to visualize single-cell genomic data. It is also known for its increased speed of computation and less stochastic in nature when compared to t-SNE. Figure 4 illustrates the UMAP plot of the preprocessed and unlabelled original data and Figure 5 illustrates the UMAP plot of the latent representation generated by the model trained with one hidden layer and 512 hidden units with a latent size of 10. The cells of the latter are labelled after having performed k-Means clustering with 6 clusters.

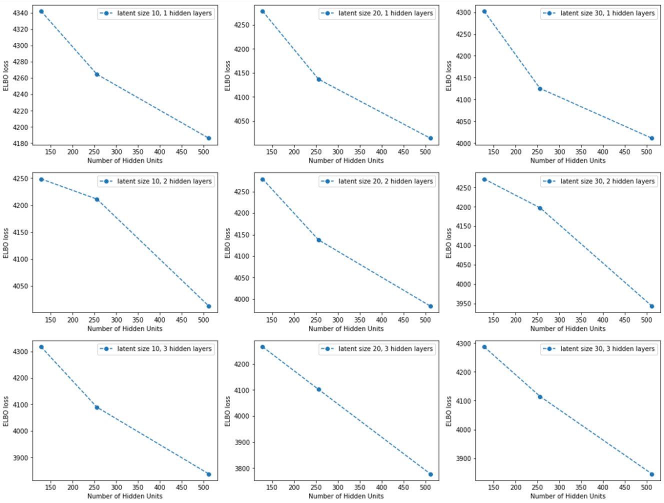

Fig. 2. ELBO loss values of different models with varying hyperparameters

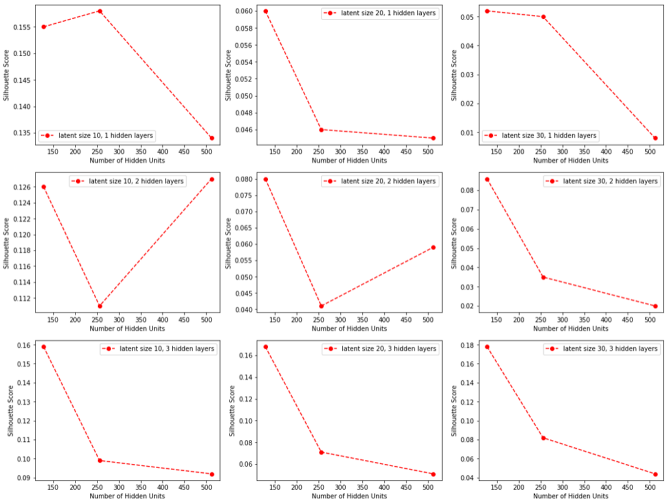

Fig. 3. Silhouette scores of different models with varying hyperparameters

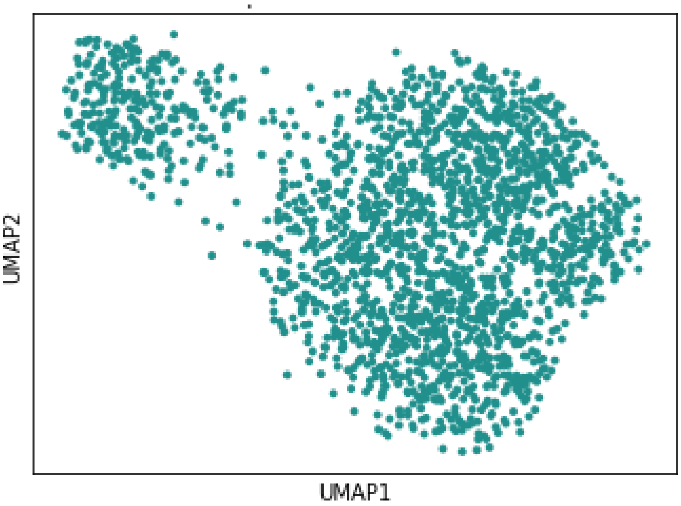

Fig. 4. UMAP plot of the original data after normalization and preprocessing

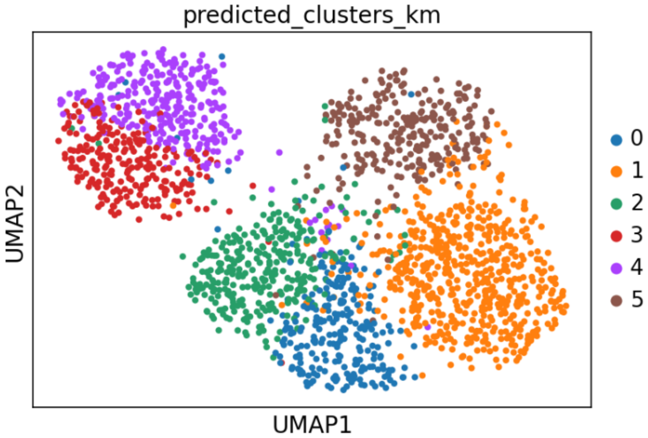

Fig. 5. UMAP plot of the k-Means clustering results of the latent space of the model trained with 1 hidden layer, latent size 10 and 512 hidden units

Differential expression and cell-type labelling

The aim of differential expression analysis is to characterize the heterogeneity of gene expression data by identifying marker genes that are differentially expressed among biological subpopulations. In the context of single-cell genomics, it is possible to determine the genes that show high or low expression for each cluster with respect to all other clusters. Following the identification of marker genes, a functional analysis would help in labelling the clusters with biologically relevant cell types. Statistical hypothesis tests like t-test and Wilcoxon-test are generally performed to estimate the confidence intervals of differential expression of genes.

In the case of the gene expression of single-cell embryoid bodies, a Seurat-based analysis was previously done on the dataset which lead to labelling the clustered data into six major progenitor cell types based on differentially expressed marker genes, as depicted in Supplementary Table 3.

These predicted progenitor marker genes are enriched in their respective cell types which helps in their development and differentiation into specific cell types. For example, stromal cells were shown to have enriched expression levels of collagen genes such as *COL3A1*, *COL5A1*, *COL5A2*, *COL1A1*, and *COL1A2*.

We aimed to reproduce these results by performing a Bayesian differential expression analysis using scVI's built-in framework, that evaluates a Bayes factor indicating which of the null and alternate hypotheses is more likely. By analysing the functions of first five out of the top twenty highly expressed genes in each cluster, we were able to manually annotate all the clusters with appropriate progenitor cell types. As depiceted in Table 2, Cluster 0 was annotated as stromal cells. The highly expressed gene (e.g., *COL1A2* and *COL3A1*) products in the cluster produce several collagen proteins, which are responsible for strengthening and supporting many tissues in the body and are often highly expressed in connective tissues. Cluster 1 was annotated as muscle cells as the genes enriched in this cluster are mostly involved in the conversion of energy from molecules like lactate (*LDHA*) and through glycolysis (*TPI1*). The highly expressed genes in cluster 2 are *MCOLN3*, *APLNR* and *CEP104*, which are mainly responsible for membrane trafficking regulation, development of blood vessels formation, heart morphogenesis and fluid direction, respectively. Hence cluster 2 was annotated as endothelial cells. *TMEM80*, the highly expressed gene in cluster 3 is found to be more involved in non-motile cilium assembly that drive fluid flow and produce signaling gradients in the brain. The proteins of another highly expressed gene from cluster 3, *MAP2* family are suspected to be involved in microtubule assembly, which is an important step in neurogenesis. Hence, cluster 3 is annotated as neural cells. Cluster 4 is annotated as epithelial cells as the genes expressed in this cluster take part in hormone inhibition, gene regulation, organogenesis and primary catabolism (e.g., *FST*, *SOX11*, *BCAT1*). Cluster 5 was labelled as liver cells as the highly expressed genes of this cluster encode for low and high density lipoprotein particles and are primarily expressed in liver. Supplementary Table 2 provides a detailed functional analysis of first five highly expressed genes in each cluster.

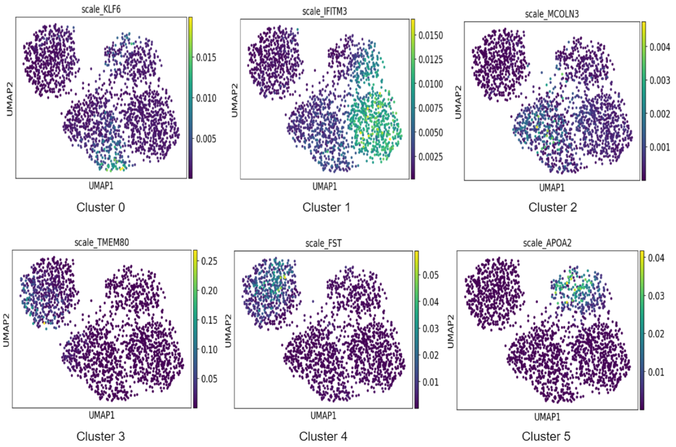

Fig. 6. Differential expression of genes that show enriched expression in each cluster; The Bayes factor associated with the genes with respect to each cluster from 0 to 5 is as follows: 2.57, 3.40, 1.87, 2.98, 2.52, 4.87

Figure 6 illustrates the expression levels of the genes with their associated Bayes factor that show highest expression in their respective cluster. Figure 7 shows the cluster-wise plots of the top 20 highly expressed genes ranked by their Bayes factors.

METHODS

Datasets

The day 8 EB datasets were obtained from Gene Expression Omnibus (GEO), which were submitted by *Han, Xiaoping et al*, whose work focused on mapping human pluripotent stem cell differentiation pathways. This study had performed dimensionality reduction, clustering and differential expression in Seurat, a package in R designed for single-cell analysis. Six major cell types of progenitor cells - muscle cells, neural cells, stromal cells, endothelial cells, epithelial cells and liver cells were identified in day 8 EB cells based on marker genes that are responsible for the development of the progenitor cells to specialized

Cluster	Top 5 highly expressed genes	Predicted Cell Type
0	*KLF6*, *FOS*, *COL1A2*, *FOSB*, *COL3A1*	Stromal Cells
1	*IFITM3*, *LDHA*, *SKP1*, *HAND1*, *TPI1*	Muscle Cells
2	*MCOLN3*, *APLNR*, *CEP104*, *HEATR1*, *ERLIN2*	Endothelial Cells
3	*TMEM80*, *MAP2*, *CRABP1*, *SOX11*, *PTN*	Neural Cells
4	*FST*, *SOX11*, *TUBB2B*, *SOX2*, *BCAT1*	Epithelial Cells
5	*APOA2*, *APOC1*, *APOB*, *TTR*, *IGFBP6*	Liver Cells

Table 1. Predicted cell types after k-Means clustering and Bayesian Differential Expression; the clusters were manually annotated based on marker genes whose functions correspond to the development and specialisation of specific cell types.

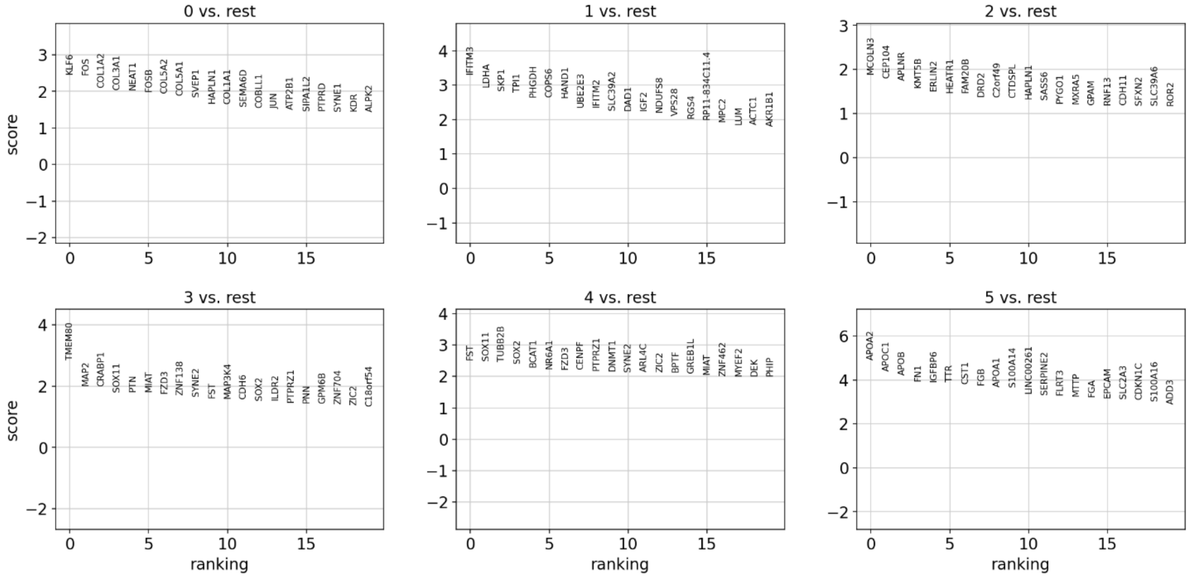

Fig. 7. Cluster-wise plots showing the top 20 highly expressed genes ranked according to their Bayes Factors

cell types. We combined datasets corresponding to four replicates of day 8 EB cells to generate a sparse gene × cell matrix that contained gene counts for 1953 cells. The batch and clustering annotations of the data were not publicly available. Hence, in this analysis, we have ignored the possibility of providing scVI with the batch labels.

Preprocessing

Scanpy, a python-based single-cell analysis toolkit, was used to load and preprocess the data. The count data matrix was first normalized by cell counts and log-transformed, following which the genes were ranked based on variance and only the highly varying genes were chosen to be included in the dataset. As scVI accepts count matrices directly as input, the filtering was performed on the original count matrix after identifying highly varying genes. The final count matrix consisted of 1953 cells and 1773 gene features.

Model training and latent space generation

scVI uses PyTorch backend for the deep learning frameworks. The inbuilt VAE function in scVI package that takes the raw count matrix, latent size, number of hidden layers and hidden units as inputs were utilised for training the model, generate the latent representations and posterior probabilities of each cell. As the batch annotations were not available for the dataset, the process of batch correction in scVI which requires batch indices as an additional input was ignored. scVI considers both the batch labels s and size factors l as random variables, the modified variational lower bound of the marginalised log-likelihood is given by

$$\begin{aligned}\log p(x|s) \geq E_{q(z,l|x,s)}&[\log p(x|z,l,s)]\\ &-D_{KL}[q(z|x,s)||p(z)]\\ &-D_{KL}[q(l|x,s)||p(l)]\end{aligned}$$

All 27 different models with varying hyperparameters were trained for 500 epochs with default values of learning and neural network dropout rates. The likelihood model was chosen as a ZINB distribution. The optimization algorithm used by scVI is Adam stochastic gradient descent.

Clustering and Visualisation

Scikit-learn, a python-based framework for implementing machine learning algorithms was used to perform clustering with Gaussian Mixture Models and k-Means algorithm on the latent representations generated by scVI. To evaluate the clustering performance, silhouette scores using the latent representations were calculated using scikit-learn's inbuilt algorithm. The number of groups was chosen to be six, and the number of random initializations was 100 for both methods. For k-Means, the maximum number of iterations was specified as 1000. Scanpy's procedure was used for latent visualisations with UMAP plots. It first computes a nearest-neighbour matrix, following which the 2-dimensional components for UMAP are generated.

Differential Expression

Bayesian differential expression was performed using scVI. It uses the expected mean frequency matrix to perform a Bayesian hypothesis testing by framing two mutually exclusive hypotheses that specify which cell among a pair of cells has a higher expression level of a gene. In each cluster, genes were ranked according to their Bayes factors, which quantifies the support for one hypothesis against the other. UMAP plots were used to visualize and compare the expression levels of genes that show enriched expression in the respective groups.

CONCLUSION

In this project, we highlight the scope and prospective applications of unsupervised deep learning techniques in the area of single-cell genomics, which is a rapidly evolving field. Precisely, we explore the capability of variational autoencoders to decompose single-cell gene expression data of human embryoid bodies and to make plausible biological inferences. We show that it is not a trivial task to choose the best set of hyperparameters for training the model and that lower loss values do not always guarantee the best model. This is evident while comparing the silhouette scores of multiple models with varying hyperparameters. Complex models with higher number of hidden layers and hidden units were shown to fit the training data well but have lower silhouette scores when compared to simpler models. Complex models can lead the algorithm to learn complex nonlinear function approximations with lower training ELBO loss values which might result in overfitting of the data. We also explore clustering algorithms like Gaussian Mixture Models and k-Means and visualisation techniques such as UMAP. The silhouette scores were lower in the case of GMMs as they are probabilistic models that are not based on distances between data points. By performing a Bayesian differential expression analysis on the expected mean frequency matrix, we were able to identify the progenitor cell types of six clusters by manual annotation. We report the highly expressed genes in each cluster, followed by a functional analysis of the predicted genes. However, a deeper understanding of the collective behaviour of genes is required to provide insights into their molecular mechanisms that lead to the specialisation of particular cell types. Additionally, it is important to note that the major progenitor cell types can be further grouped into biologically relevant subclusters that contribute to the heterogeneity within individual clusters.

In the future, we wish to use computational techniques to investigate further the developmental processes that can provide insights into the exact mechanisms behind the proliferation and differentiation from a single-cell zygote to highly specialized cell types with diverse functions. One approach would be to profile the expression levels of embryonic stem cells at several instances of time and perform a time series analysis to understand the time-dependent changes in expression levels that influence their developmental trajectories. We look forward to addressing the shortcomings and drawbacks of variational autoencoders such as stochasticity, nonconvex optimization and parameter explosion.

DISCUSSION

In recent years, with the development of a range of analysis tools, scRNA-seq data analysis has significantly advanced our knowledge of biological systems. All standard scRNA-seq analysis pipelines include preprocessing steps such as normalization, removal of poorly varying genes and cells having high expression levels of mitochondrial genes. However, preprocessed data might still have unwanted technical and biological noise factors such as batch, dropout and cell cycle effects, which must be corrected to single out particular biological signals of interest. Existing toolkits accomplish cell cycle effect correction by a simple linear regression against cell cycle score. The batch effect is the systematic difference in gene expression levels between different batches of the dataset. Several techniques have been formulated to eliminate batch effects. It is also important to prevent the loss of biological information while preprocessing.

Tools that use embedded deep learning workflow include effective decomposition, data fitting and denoising procedures to perform downstream analyses such as clustering and differential expression within the framework of the model. Models trained using toolkits like scVI learn a cell-specific scaling factor and latent encodings as hidden variables, with the objective of maximizing the likelihood of the data while also taking batch and dropout effects into account. These tools can be further extended to merge multiple datasets from a given tissue. Automated cell type annotation tools can also be integrated to annotate cell types after clustering the scRNA-seq data.

The next stage of development will be integrating multi-omics data. Integrating different levels of genomic and proteomic information can be important to fully decipher the changes in stem cells during proliferation and differentiation which can help to identify subsets of rare cells with specific phenotypes or certain molecules with valuable information on stem cell lineage. It will also be necessary to link this information with clinical and other metadata, such as the tissue origin, flow cytometry data or with the sample origin and disease diagnosed.

With the advancements in single-cell RNA-sequencing technologies and their improved capacity to capture the expression levels of a large number of cells, we anticipate that there will be a massive explosion in the amount of single-cell data in the near future. Hence, through this project, we emphasize the importance of deep learning techniques for single-cell genomic analysis due to their robust frameworks, scalability and their ability to make useful biological inferences.

SUPPLEMENTARY INFORMATION

The supplementary information contains additional tables and figures that provide information about the other models that were trained apart from the one chosen for downstream analysis. The functions of multiple genes that were not highlighted in the report are tabulated in the supplementary file.

Link to supplementary information

The codes and commands that we used are available in our GitHub Repository

Link to view GitHub Repository

ACKNOWLEDGEMENTS

We are extremely grateful to Professor Karthik Raman for providing us with immense knowledge about the various aspects of Computational Systems Biology. His constant guidance encouraged and motivated us to explore a variety of research avenues.

REFERENCES

- Han X, Chen H, Huang D, et al. Mapping human pluripotent stem cell differentiation pathways using high throughput single-cell RNA-sequencing. Genome Biol. 2018;19(1):47. Published 2018 Apr 5. doi:10.1186/s13059-018-1426-0
- Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nat Methods. 2018;15(12):1053-1058. doi:10.1038/s41592-018-0229-2
- Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., Theis, F. J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. Nature communications, 10(1), 390. https://doi.org/10.1038/s41467-018-07931-2
- Chen, Sisi, et al. ”Dissecting heterogeneous cell-populations across signaling and disease conditions with PopAlign.” bioRxiv (2018): 421354.
- Yilmaz, Atilgan, et al. ”Haploid human embryonic stem cells: half the genome, double the value.” Cell Stem Cell 19.5 (2016): 569-572.
- Young, Richard A. ”Control of the embryonic stem cell state.” Cell 144.6 (2011): 940-954.
- Jaenisch, Rudolf, and Richard Young. ”Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming.” Cell 132.4 (2008): 567-582.
- Baker, Monya. ”Embryoid bodies get organized.” Nature Reports Stem Cells (2008): 1-1.
- Nguyen, Quan H., et al. ”Single-cell RNA-seq of human induced pluripotent stem cells reveals cellular heterogeneity and cell state transitions between subpopulations.” Genome research 28.7 (2018): 1053-1066.
- Amodio, M., van Dijk, D., Srinivasan, K. et al. Exploring single-cell data with deep multitasking neural networks. Nat Methods 16, 1139–1145 (2019). https://doi.org/10.1038/s41592-019-0576-7
- Cao, Yinghao Wang, Xiaoyue Peng, Gongxin. (2019). SCSA: a cell type annotation tool for single-cell RNA-seq data. 10.1101/2019.12.22.886481.
- Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol. 2019;15(6):e8746. Published 2019 Jun 19. doi:10.15252/msb.20188746
- Samir, Jerome Rizzetto, Simone Gupta, Money Luciani, Fabio. (2020). Exploring and analysing single cell multi-omics data with VDJView. BMC Medical Genomics. 13. 10.1186/s12920-020-0696-z.
- Li, X., Wang, K., Lyu, Y. et al. Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. Nat Commun 11, 2338 (2020). https://doi.org/10.1038/s41467-020-15851-3
- Stuart T, Butler A, Hoffman P, et al. Comprehensive Integration of Single-Cell Data. Cell. 2019;177(7):1888-1902.e21. doi:10.1016/j.cell.2019.05.031
- Wolf, F., Angerer, P. Theis, F. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol 19, 15 (2018). https://doi.org/10.1186/s13059-017-1382-0
- McCarthy DJ, Campbell KR, Lun AT, Wills QF. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinformatics. 2017;33(8):1179-1186. doi:10.1093/bioinformatics/btw777