Variational Auto Encoder Approach To Find Deferentially Expressed Genes

نهج التشفير التلقائي المتغير على للعثور الجينات المعبر عنها بشكل تفضيلي

by

NABIL RAHIMAN

Dissertation submitted in partial fulfilment

of the requirement for the degree of

MSc INFORMATICS

at

The British University in Dubai

May 2022

DECLARATION

I warrant that the content of this research is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.

I understand that a copy of my research will be deposited in the University Library for permanent retention.

I hereby agree that the material mentioned above for which I am an author and copyright holder may be copied and distributed by The British University in Dubai for the purposes of research, private study, or education and that The British University in Dubai may recover from purchasers the costs incurred in such copying and distribution, where appropriate.

I understand that The British University in Dubai may make a digital copy available in the institutional repository.

I understand that I may apply to the University to retain the right to withhold or to restrict access to my thesis for a period which shall not normally exceed four calendar years from the congregation at which the degree is conferred, the length of the period to be specified in the application, together with the precise reasons for making that application.

Signature of the student

COPYRIGHT AND INFORMATION TO USERS

ABSTRACT

A study of differentially expressed genes across different cell types will help in identifying cell-specific responses to treatments or diseases. Recent advances in single-cell technology enable an analysis of thousands of cells which brought lots of computational challenges in terms of noise in the data sets and required computational power to handle the big data. In recent years it has been found that the deep learning model is being used as a biological model for single-cell analysis. Using state-of-the-art techniques in deep learning successfully extracts non-linear feature set from single-cell data and is used for various downstream analysis.

Recently, deep learning models such as Autoencoder (AE) and Variational Autoencoder (VAE) models are being used to capture hidden patterns from single-cell gene expression data. In this paper, I proposed a framework that is based on a variational autoencoder called BiDiffVAE (Bi-directional Differential Variational Autoencoder) to extract differently expressed genes. The proposed method makes use of cluster distribution on every latent space and merged weights in the decoder to assign genes to a cluster. My results discovered new sets of genes that were not shown using state-of-the-art techniques and can properly rank the top genes based on their significance in making clustering.

ملخص

تتيح التطورات الحديثة في تقنية الخلية الواحدة تحليل آلاف الخلايا التي جلبت الكثير من التحديات الحسابية من حيث الضوضاء في مجموعات البيانات والقدرة الحسابية. في السنوات الأخيرة، تم العثور على نموذج قائم على التعلم العميق يستخدم كنموذج بيولوجي لتحليل خلية مفردة ولتحليل استجابة معالجة السجادة. يؤدي استخدام أحدث التقنيات في التعلم العميق إلى استخلاص مجموعة الميزات غير الخطية بنجاح من بيانات الخلية المفردة واستخدام العديد من التحليلات النهائية. أقترح في هذه الورقة طريقة تعلم عميق لاستخراج الجينات المعبر عنها بشكل مختلف دون استخدام أي أداة إحصائية. تقوم الأداة المقترحة بتعيين الجينات إلى الكتلة وتعطي ترتيبًا يشير إلى مدى تميزها عن مجموعة أخرى. اكتشفت نتائجي مجموعات جديدة من الجينات لا تظهر في أحدث التقنيات المستخدمة في التعلم العميق

ACKNOWLEDGMENT

I would like to express my gratitude to my primary supervisor, Professor Sherief Abdalla, who guided me throughout this project. His feedback was very valuable in building the thesis paper

I would also like to thank my collogues at NYU, Abu Dhabi who supported me and offered a deep understanding fundamentals of genomics.

It is my privilege to thank my family for their consistent encouragement, support, and prayers throughout my master's course.

I am extremely thankful to Professor Sherief Abdalla and Prof. Khaled Shaalan for providing their valuable feedback through the courses and their research-oriented teaching methods helped in reviewing state-of-the-art techniques in the area of linguistic and deep learning.

Finally, I would like to thank God, the Almighty who has granted me the countless blessing and let me go through all difficulties and achieve the goals.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE: INTRODUCTION

All individuals' health state is controlled by their genes and their environment. Environmental changes can cause changes in genes which can lead to the development of disease. Identifying genetic variation in gene expression can help understand genetic disorders caused by mutations and targeted gene-drug discovery. A living organism is built from the smallest unit called a cell. Every cell has copies of hereditary information is encoded in a DNA molecule. DNA is responsible for the development of a living organism. Gene expression is the process of making cellular biological components based on instructions encoded in DNA. Identifying cell types in a sample collected from an organ can help researchers to understand how a particular cell type responds to disease or medicine.

Advances in single-cell sequencing technologies accelerate the research in single-cell RNA sequencing to explore the RNA transcripts at a single cell granular level. This reveals heterogeneity in single-cell types and complexity within a single cell type. Gene expression analysis is a popular method (Bondoc et al., 2021) to study living organisms at the single-cell level.

Marker genes of cell type are genes that are overexpressed in that cell type and not expressed highly in other cell types. Marker genes distinguish the cell subpopulations in given data sets. A very common step in single-cell RNA sequencing analysis is selecting marker genes and cell-type identification. Marker genes are selected concerning cluster and its selection is very critical for labeling clusters with cluster type and downstream analysis. Incorrect marker gene selection can cause incorrect downstream analysis. Though different statistical tools are available to find out the differentially expressed genes, the deep learning model is found to be the state-of-the-art technology to capture the differentially expressed genes. One of the major challenges in handling a large volume of single-cell high-throughput sequencing data is noise in the data. Very common technical noises are sparsity and variations in the data that cause other than biological factors. Deep learning models such as autoencoder (AE) and variational autoencoder (VAE) are state-of-the-art techniques for solving the technical noise in the data.

Since variational autoencoder (VAE) has an inbuilt feature to remove technical noise, in this paper I propose a method called BiDiffVAE (Bi-directional Differential Variational

Autoencoder) that uses VAE to find the differentially expressed genes. BiDiffVAE is an extended version of DiffVAE (Bica et al., 2020). In DiffVAE first encoded values are used for labeling latent dimension by a cluster type. To do this, it computes all cell distribution at least cell value a standard deviation from the mean in latent dimension k. The latent dimension will be labeled by the cluster type whose distribution is the highest in the same latent space. Once cluster type is identified merged weights in the decoding layer are used to get to differentially expressed genes. This method has a side effect. It didn't consider all latent space together while labeling the genes.

In the proposed model, cluster distribution at each latent space is also considered. Every gene compute cluster weight associated with a particular cluster by computing the sum of each merged weight in the decoding layer multiplied by the corresponding cluster distribution on latent space. If found in N clusters, cluster weights computation will be repeated for all N clusters by a particular gene. Gene will be labeled by cluster type with higher cluster weights.

1.1 Problem Statement

As stated above model based on VAE is capable of modeling biological interpretation, hence I explore a method to find differentially expressed genes based on VAE.

1.2 Research Question

- How to include the cluster wights in finding differentially expressed genes?
- How does including the cluster distribution improve the performance?

1.3 Contribution

In this paper, I propose BiDiffVAE, an extension to the DiffVAE (Bica et al., 2020). I evaluated my proposed approach on real-world data set and show that my proposed approach outperforms the state-of-the-art in terms of gene coverage and the ability to rank genes concerning the gene that is differentially expressed.

1.4 Scope

Finding top driving genes is very critical in cell type identification and gene set enrichment analysis (GSEA). GSEA(Subramanian et al., 2005a) is a method to find overrepresented pathways in the gene list.

1.5 Organization of Thesis

The thesis starts with an overview of this paper. I tried my best to give a background on the fundamentals of biology to get an understanding of genomics and machine learning to address both audiences who has either no biology background or no computer science background. In the related work section, reviewed papers based on variational autoencoder which used VAE as a biological model for gene expression. After the study of related work, discussed the methodology finally have a discussion section to conclude for finding and future scope.

2 CHAPTER TWO: BACKGROUND

A cell is the smallest unit in a living organism that lives on its own and makeup all living organisms. Two types of living organisms are based on the number of cells in an organism.

- Single cell: Examples are bacteria and yeast
- Multicell: Animals and plants

All cell has three basic parts

- Cell membrane: Outermost layer that separates the cell interior of cells from the outside environment and protects the cells from their environment.
- Cytoplasm: Inside a cell membrane contains a liquid material called Cytoplasm. It is a medium for chemical reactions. Many cellar operations occur in the cytoplasm.
- Nucleus: Found inside the nucleus.

Based on structural differences, organisms (Prokaryotes vs Eukaryotes) are grouped into two categories:

- Eukaryotes: These are organism which doesn't have nucleolus and mitochondrial genes
- Prokaryotes: DNA is found inside the nucleolus.

Cell together forms tissue and tissue forms together with an organ. The order of things is shown in figure 1.

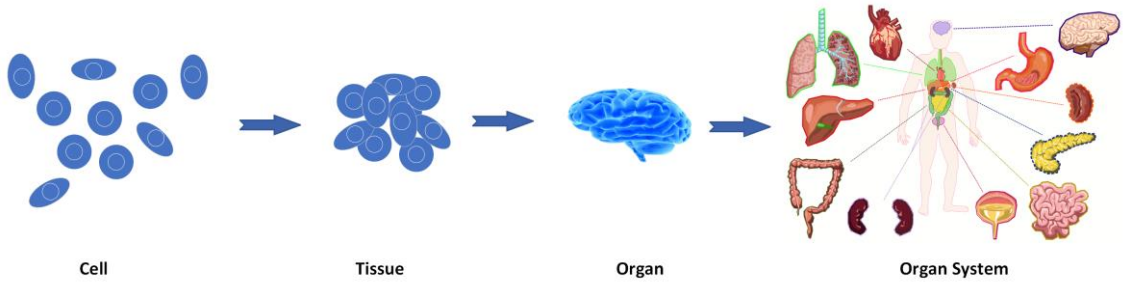


Figure 1: Levels of organization in a living organism

A human body consists of trillions of cells (Bianconi et al., 2013). A cell contains hereditary materials and cells can make copies of themselves. Hereditarily material is called DNA (Deoxyribose Nucleic Acid), carrying instructions for the development, and functioning growth of the living organism. DNA is like a ladder structure and is made up of 4 chemical bases adenine (A), guanine (G), cytosine (C), and thymine (T). A long strand of DNA is arranged as a tightly coiled structure called a chromosome. In humans, there are 23 pairs of chromosomes. One of the pairs is inherited from the mother and the other from the father.

All DNA in one cell is called a genome. A human genome contains 3 billion bases (ACGT) pairs. A gene is A portion of the long DNA strand.

The majority of DNA is found inside the nucleus. All cells contain the same DNA sequence, based on the cell type certain genes will be turned on or off, which determines what kind of proteins needs to be generated for the certain type of cell.

Table 1: Functions of proteins.

Function	Description
Antibody	Antibodies bind to specific foreign particles, such as viruses and bacteria, to help protect the body.
Enzyme	Enzymes carry out almost all the thousands of chemical reactions that take place in cells. They also assist with the formation of new molecules by reading the genetic information stored in DNA.
Messenger	Messenger proteins, such as some types of hormones, transmit signals to coordinate biological processes between different cells, tissues, and organs
Structural component	These proteins provide structure and support for cells. On a larger scale, they also allow the body to move.
Transport/storage	These proteins bind and carry atoms and small molecules within cells and throughout the body.

("What are proteins and what do they do? MedlinePlus Genetics," n.d.)

Gene Expressions (Proteins Synthesis):

The process of making cell functional products (proteins) from instructions coded in a gene is called gene expression. Gene Expression is carried out in a two-stage process, transcription, and translation. In the transcription stage, mRNA will be created using DNA as a template. mRNA is a type of RNA which is a kind of single-stranded DNA.

RNA polymerize is the process that unwinds the DNA to create mRNA. There are regions in the DNA where polymerize starts and ends. **mRNA** contains an Uracil molecule(U)

instead of thymine(T). The initial stage of mRNA is called pre-mRNA which contains both intron and exon. Exon and intron are two regions in RNA strand, where the exon region codes for the protein and the intron section is called the non-coding region. During the process of intron splicing, the intron regions will be removed from pre-mRNA and produce mature RNA (mRNA).

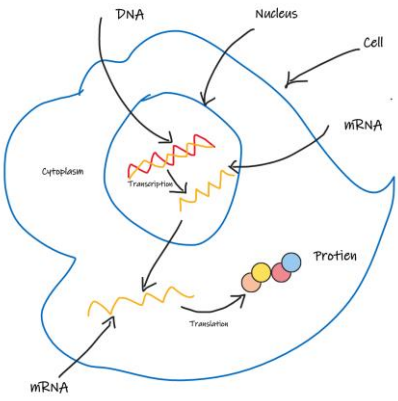


Figure 2: An overview of the flow of information from DNA to protein

mRNA leaves the nucleus to start the translation process. Translation begins at the cytoplasm. Three letter code in the mRNA is called a codon which codes for 64 different combinations of codons, that code for 20 different amino acids (“The Information in DNA Determines Cellular Function via Translation | Learn Science at Scitable,” n.d.). 61 codons encode amino acids and the remaining three encode stop signals. There is special amino also known as start codon (AUG). Amino acids are joined together to make polypeptides. Polypeptide foldup together makes proteins.

Second Nucleotide												Third Nucleotide
	U			C			A			G		
U	UUU	Phe		UCU	Ser		UAU	Tyr		UGU	Cys	U
	UUC			UCC			UAC			UGC		C
	UUA	Leu		UCA			UAA	STOP		UGA	STOP	A
	UUG			UCG			UAG			UGG	Trp	G
C	CUU	Leu		CCU	Pro		CAU	His		CGU	Arg	U
	CUC			CCC			CAC			CGC		C
	CUA			CCA			CAA	Gln		CGA		A
	CUG			CCG			CAG			CGG		G
A	AUU	Ile		ACU	Thr		AAU	Asn		AGU	Ser	U
	AUC			ACC			AAC			AGC		C
	AUA			ACA			AAA	Lys		AGA	Arg	A
	AUG	Met		ACG			AAG			AGG		G
G	GUU	Val		GCU	Ala		GAU	Asp		GGU	Gly	U
	GUC			GCC			GAC			GGC		C
	GUA			GCA			GAA	Glu		GGA		A
	GUG			GCG			GAG			GGG		G

Figure 3: Codon encoded amino acids

Transcriptomics

Transcriptomics is the study of transcriptomes (RNA transcripts). Following are two main techniques to measure RNA

- Micro-arrays: Profiles predefined transcripts/genes
- RNA-Seq: Sequencing of the whole transcriptome

RNA-Sequencing is a method to measure gene expression at the whole transcriptome level. RNA sequencing can be done in two says

- Bulk RNA sequencing
- Single-cell RNA sequencing

In bulk RNA sequencing average gene expression is measured at the population level. Whereas in single-cell RNA sequence(scRNA-seq) gene expression is measured at the single-cell level. Cell to cell gene variability could be measured using the scRNA-seq method. Single-cell technology was introduced in 2009 starting by analyzing a single cell and the scale of handling cells increased exponentially over the past decades (Svensson et al., 2018).

Sequencing technology differs based on the following characteristics (Chen et al., 2019).

- Cell isolation
- Cell lysis
- Reverse transcription
- Amplification
- Transcript coverage
- Strand specificity
- Molecular tags that can be applied to detect and quantify the availability of the unique transcript

Machine Learning

Machine learning is the study of algorithms that can improve automatically through experience by making use of the data. There are two categories of machine learning, supervised and unsupervised. The supervised learning model used labeled data for the

training. It solves two types of problems, classification, and regression. Classification is a supervised learning method that classifies the data into categories. Regression is another kind of supervised learning method that finds the relationship between dependent and independed variables.

The unsupervised learning model uses label data for the training. In scRNA-seq analysis, an unsupervised learning model is typically used for clustering and reducing the dimensionality of gene expression data. Clustering is a kind of classification problem, but here the cells are classified based on the cell similarity measurement.

Deep learning

Machine learning is always followed by feature engineering tasks. Feature extraction is sometimes very difficult if we have complex data. The deep learning model extracts hidden features on its own, and hence not required a feature engineering task. A deep learning model is constructed from many artificial neural networks with many neural network layers. Neurons on every layer are connected to other neurons in the next layers as given in figure 4.

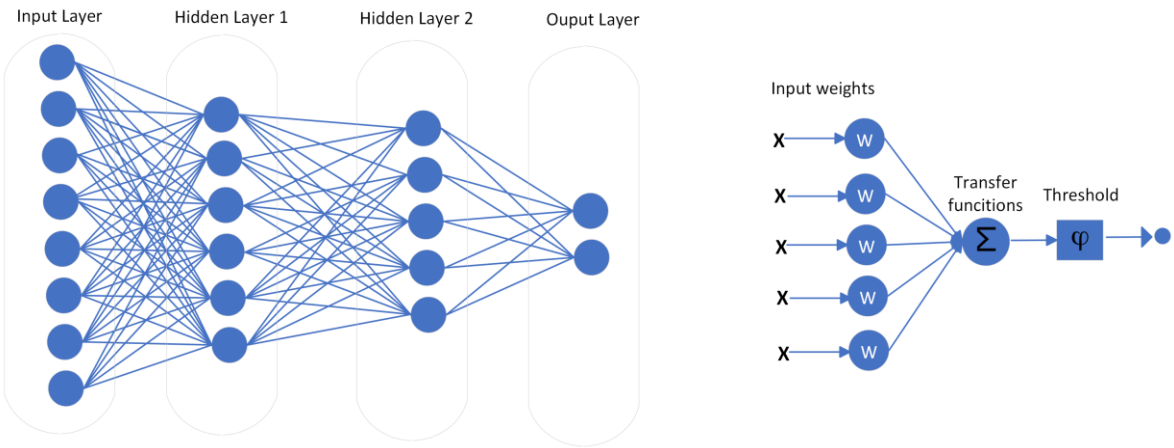


Figure 4: A typical artificial neural network

In this paper, we are mainly focused variational autoencoder model (VAE) which is a variant of the autoencoder model. The basic autoencoder (AE) model has an encoder and decoder neural network which has symmetrical in shape as in figure 5.

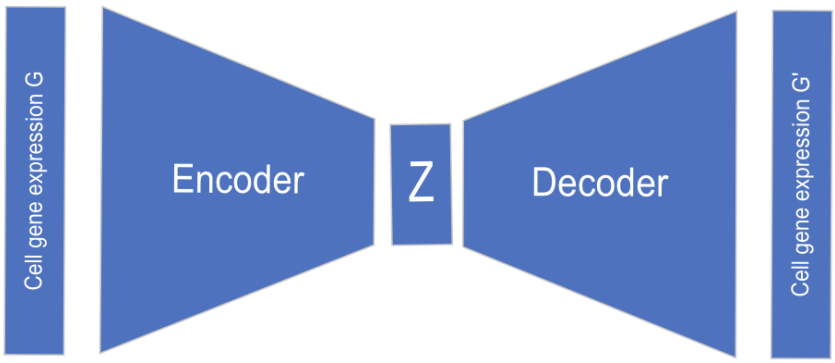


Figure 5: Autoencoder Model (AE)

AE creates a lower dimension space of original data called latent space or bottleneck. The decoding layer is responsible for reconstructing input data from latent space. The objective of this model is to minimize the reconstruction error. Autoencoder is widely used in clustering, denoising, and dimensionality reductions (Eraslan et al., 2019).

The main difference between Variational autoencoder (VAE) and autoencoder (AE), VAE model is to approximate latent distribution to be a Gaussian distribution. The encoder maps each input to a mean vector and variance in the latent space. The latent variable is sample data from a normal distribution which is scaled by the standard deviation and mean vector computed. This makes it model a generative model

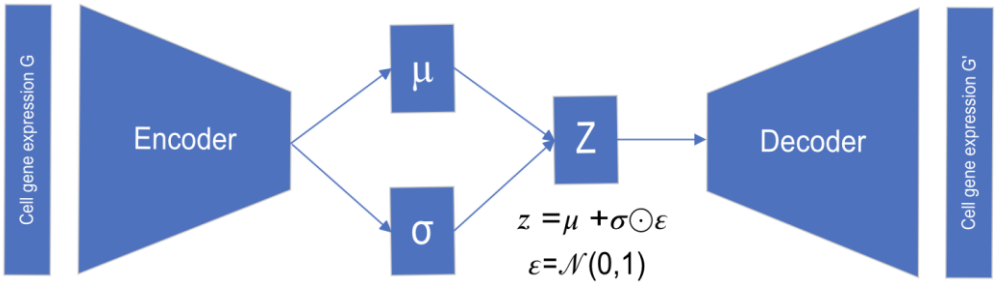


Figure 6: Variational Autoencoder(VAE)

Differential gene expression is important to understand the biological differences between healthy and diseased states. RNA-seq is a popular technique to quantify gene expression between conditions. The following section is covering the single cell RNAseq pipeline.

Single-cell RNA-seq analysis pipelines

A typical single-cell RNA-seq analysis pipeline consists of preprocessing, clustering, visualization, cluster annotation, and different downstream analysis. Raw data from the RNA sequencing machine are processed and aligned to produce a gene count matrix. Preprocessing steps take this count matrix for quality control, normalization, data correction, feature selection, and dimensionality reduction(Luecken and Theis, 2019). In downstream analysis main task is to cluster and cluster annotation or cell type identification. Finding the differentially expressed gene in each cluster will help in identifying marker genes. Marker genes are signature genes for specific cell types.

Common Computational Challenges

Large volume for single-cell high-throughput sequencing data and multiple characteristics of single-cell sequencing data leads to computational challenges. The single cell isolation technique is a method to extract transcriptome data from individual cells. Though different protocols are available to extract transcriptome data (Hwang et al., 2018), transcriptome data is usually generally accompanied by higher noise and dropout rates. Due to cell heterogeneity in gene expression data, finding rare cell types is very difficult(Fang et al., 2021). Another problem is that scRNA-seq suffers excessive zero (Fang et al., 2021). Some zero counts are true zeroes and others are not. Imputation is a method to address the increased sparsity observed in scRNA-seq data.

Integration of data from multiple sources and analysis of single-cell RNA sequencing (scRNA-seq) remains challenging due to the variation other than biological factors. The batch effect (Liu et al., 2020) can also cause differences in gene expression due to non-biological factors. Batch effects can cause false clustering which leads to incorrect downstream analysis. This incorrect clustering will end up in the wrong conclusion. Later in

the section of this paper, I go over the state-of-the-art techniques that address these problems. A literature review is conducted mainly that focus on the deep learning approach.

The typical feature size of scRNA-seq is around 20,000 and computing cell difference takes a lot of computational power and is difficult to visualize. The expression of many genes correlated; hence we don't require all the genes to cluster similar cells. By lowering the dimensions, it keeps the most important features in the data sets. Hence it removes duplicate features from the data sets. Low dimensional data should hold the properties of higher-dimensional data. The dimensionality reduction method reduces the number of features in data sets keeping the distribution of the original data. Dimensionality reduction methods help the researcher in various ways. It helps in visualizing high-dimensional data and minimizes the computation time and resources it consumes. If the data is a manifold space other than Euclidian space, then the non-linear dimensionality reduction method is required to bring down the higher dismissions

Limited transferability is another problem mainly because of the batch effect and limited access to the data. This way the knowledge learned from one dataset cannot be easily transferred to benefit the modeling of another dataset. For reference data construction, integration methods required access to the data that has limited access. Small data sets are not adequate for training a model, it is required a larger amount of data. In such cases, the transfer learning model is useful when you have small data sets.

A very common step in single-cell RNA sequencing analysis is selecting marker genes. Marker genes are selected with respect to clusters. Marker genes selection is very critical for annotating clusters and downstream analysis. Marker genes distinguish the cell subpopulations in given data sets. There are exist many computational methods available that range from statistical to machine learning techniques. The statistical model usually has limited transferability; hence it is required very times to run the entire analysis pipeline. If the model support transferable learning, only during the training phase does it have to go over the pipeline. Once is learned the model; it may be needed to run the pipeline again.

3 CHAPTER 3: RELATED WORK

Recent studies modified versions of the original Autoencoder (AE) and Variational autoencoder (VAE) framework for modeling single-cell RNA sequencing analysis. In this section, I explain the motivation behind using variational autoencoder for finding differential expressed genes. VAE is found to be powerful in exploring hidden biological patterns in gene expression. The low-dimensional latent space layer in VAE enforces the encoder to learn only the essential latent representations and the decoding procedure ignores non-essential sources of variations of the expression data. I search the latest research papers based on VAE/AE as a computational tool using the following inclusion/exclusion principle.

Table 2: Inclusion Exclusion Principle

Es32w1q	Exclusion Principle
("Autoencoder" or "variational autoencoder") and ("single-cell RNA")	Excluded the papers if not using scRNA seq data as input data
Included the papers from nature.com, pubmed.ncbi.nlm.nih.gov	Ignore the data is not a single cell RNA sequencing data
Included in the literature if common techniques discussed in most of the paper	

Initialize literature all are based on tools that are primarily used autoencoder (AE) for denoising, dimensionality reduction, and cluster labeling. Deepimpute(Arisdakessian et al., 2019)is a standard autoencoder based on an artificial neural network with a dropout layer included. A dropout layer is included to avoid overfitting the data. DCA(Eraslan et al., 2019) used autoencoder as a computational framework for performing mainly data imputation. Since it is an artificial neural network, it can capture non-linearity in the data, encoding part of the model compresses the data into low dimensional space it can be used as a tool for dimensionality reduction. DCA model replaces the conventional mean square error (MSE) loss function with a zero-inflated negative binomial (ZINB) model-based loss function. scScope(Deng et al., 2019) is another tool based on autoencoder which uses a recurrent neural network where the output is connected back to the encoder to improve imputation performance iteratively. scDeepCluster (Tian et al., 2019) follows DCA which uses ZINB

model-based loss function, and clustering loss function (KL-divergence) is applied. In Sparsely Connected Autoencoder (SCA) encoding and decoding module consisted of a single sparse layer with connections based on known biological relationships. MARS(Brbić et al., 2020) is a two-stage model, in the first stage weights were assigned with a deep autoencoder network and then perform learning cell-type landmarking after removing the decoder from the autoencoder.

So far, the discussion has reviewed the tools which are based on autoencoder models. From here onwards, the literature will focus mainly variational autoencoder-based model. VEGA (Seninge et al., 2021) is a generative model, based on VAE for inferring the biological model from its latent representation. VEGA's latent space could group the control and treated cells separately in response to different perturbations. In VEGA each latent dimension defines as a gene module variable (GMV) which could be cell type, pathway, or gene regulator. Pathways are biologically related genes. GSEA(Subramanian et al., 2005b)gene set enrich analysis is used to search for pathways that were statistically significant with target phenotypes. Example target phenotypes are diseased vs healthy. Since VEGA's latent dimension define a pathway, this model is used for enrichment analysis. In scETM(Zhao et al., 2021) variation autoencoder is used for constructing the embedded topic model. In typical VAE, encoder and decoder are symmetrical in structure. In scTEM encoder is constructed from the nonlinear neural network, but the decoder is constructed using a linear decoder using matrix tri-factorization. scTEM outperforms when tested with unseen data with zero-transfer learning performance. In resVAE(Lukassen et al., 2020) Restricted latent variational autoencoder (Lukassen et al., 2020) decoder is modeled using a sparsed weighted matrix. Spared weighted matrix is constructed based on prior knowledge about pathways or gene sets. SCVI (Ding et al., 2018) takes raw gene expression count and assumes data distribution is kind of (Poison, ZINB, NB). scVI is a bayesian variational autoencoder that accounts for batch-specific variation and it applies the ZINB-loss function to optimize the performance. SCVIS(Ding et al., 2018) uses the student's t-distributions instead of the loss function mean square error. trVAE(Lotfollahi et al., 2020) is an improved variational autoencoder (VAE) structure to pre-train the multiple datasets simultaneously. It is built upon a conditional variational autoencoder and uses the regularization method maximum mean discrepancy (MMD) in the decoding path. SCANVI(Xu et al., 2021) is a scVI based framework where it uses the annotated label to improve the cell type assignment.

scArches(Lotfollahi et al., 2021) maps the gene expression onto reference datasets for data integration and identification of cell types. Existing approaches like the Seurat (Hao et al., 2021) platform allow for integration of data but require that users run the complete pipeline on new datasets which requires excessive computational resources and time. scArches uses a transfer learning approach to transfer the knowledge from a pre-trained model to user-specific data. The pre-trained model will be useful annotate unseen data without any kind of delay. scGEN (Lotfollahi et al., 2019) predicts the perturbation response of unseen species or cell types. It used vector arithmetic to compute the difference in the perturbation response in latent space. LDVAE(Svensson et al., 2020) is a linear-decoded variational autoencoder where it uses a linear model in the decoding layer. DeepSEM(Shu et al., 2021) is a variational autoencoder-based model where it uses the adjacency matrix of the GRN in both encoder and decoder. In general, a single cell feature from gene expression data will be input into the model, whereas in DeepSEM single gene feature is used as input the for model. Hence weight in the neural network will be shared across all genes. DeepTCR(Sidhom et al., 2021) implemented a variational autoencoder. The encoding layer is reconstructed by use of deconvolutional and fully connected layers. The scDHA core modules were constructed using two stacked encoders, the first encoder is a vanilla autoencoder, and the second one variational autoencoder.

Principle component analysis (PCA) is a popular unsupervised linear dimensionality reduction method to characterize cell types and cell states. PCA finds linear projection of the high dimensional data where the variance of projected data is maximized. Euclid distance will not work with manifold data, as the actual distance could be much larger in the manifold space. So PCA doesn't work so well for visualization as its preserves large pairwise distances. tSNE (van der Maaten and Hinton, 2008) is a non-linear dimensionality reduction method. tNSE works by maximizing the probability between two points remaining close in a low dimension space as they were in the original dimensional space. tSNE suffers a performance penalty if the data sets are too large and suffers from inter-cluster relationships. UMAP (Becht et al., 2018) is like tSNE but differs in similarity measurement. UMAP gives better computation performance. Seurat (Hao et al., 2021) is a popular tool for single-cell RNA-seq analysis. It uses both UMAP and tSNE for the visualization process. To speed the UMAP and tSNE, the Seurat package first runs PCA to a reasonable number of dimensions before transferring the data into two-dimensional space.

During data integration, there are non-biological factors that can cause changes in the data produced by different experiments. Examples of confounding factors are different processing times and different handlers. Seurat (Hao et al., 2021) introduced a package that used state an art techniques method based on the anchor which will remove the batch effect from scRNA-seq datasets. The idea is to identify the cells across samples having the same biological states. These cells will be used to correct technical differences observed in the data sets collected from different samples.

To identify differentially expressed genes, very commonly used methods are FDR (false discovery rate) and log fold changes (Kamath et al., 2022). FDR uses adjusted p-value to measure the null hypothesis. The statistical test uses two groups for testing. If we have multiple clusters, then one cluster is compared with all other clusters. Log fold changes are typically used to understand whether the genes are upregulated or downregulated. Commonly used differential expression testing software packages are Seurat (Hao et al., 2021) which is an R package, and Scanpy (Wolf et al., 2018) which is a python package.

Seurat uses the following statical test method for differential expression testing

- Wilcoxon rank sum test (default)
- Likelihood-ratio test for single-cell feature expression (McDavid et al., 2013)
- Standard AUC classifier
- Student's t-test
- Likelihood ratio test assuming an underlying negative binomial distribution. Use only for UMI-based datasets
- Likelihood ratio test assumes an underlying negative binomial distribution
- Logistic regression framework to determine differentially expressed genes.
- MAST: (Finak et al., 2015)
- DESeq2 (Love et al., 2014)

Scanpy uses the following statical test method for differential expression testing:

- t-test,
- t-test with overestimation of variance of each group,
- Wilcoxon rank sum test,
- logistic regression

Since these all tests are statistical models, it is required to rerun all single-cell analysis pipelines from the beginning. It also has a scaling issue if the number of features and input data sets is too large. Machine learning model has transfer learning capacity, hence trained model can be applied to unset input data for prediction.

DiffVAE (Bica et al., 2020) introduced a method to find differentially expressed genes in a cluster using weights associated between output layers (Gene expression reconstructed) and latent space. DiffVAE use modified vanilla VAE with batch normalization. In this paper, we propose an improved version of DiffVAE which gives an accurate measure of differential expressed gene lists ordered by their importance in constructing clusters. Later this section it is been elaborated in detail on the architecture and performance results.

CHAPTER 4: METHODOLOGY

By using the unsupervised learning model VAE-variational autoencoder, it can get differentially expressed genes that separate clusters from other cluster groups. The proposed methodology is motivated by the paper (Bica et al., 2020) that uses a variational autoencoder to find out the differentially expressed genes. But the top listed genes were not uniquely expressed in those clusters. My proposed solution gives high-quality genes that separate a cluster from other clusters in the group by their expression values. The top listed gene are ordered based on their significance in making clusters

4.1 Data sets

To do a performance comparison I use the same set of data sets from the paper (Bica et al., 2020) and followed the same sets of preprocessing as described in (Bica et al., 2020). The zebrafish data sets have gene set size 1845 which are considered to be highly variable genes and 1422 zebrafish cells.

4.2 Software Tools

To model variation autoencoder we use python3+ and the following packages for preprocessing, training, and visualization

- Sklearn
- Pandas
- Seaborn
- Keras

4.3 Preprocessing

Gene expression data is the first log normalized and then transformed it using Min-Max scaling to retain the expression values within [0, 1]. Sometimes gene expression may contain extreme values known as outliers and the performance of deep learning could be impacted by these outliers. Min-Max scaler can bring outlier closer to [0,1]. If are using raw data directly from RNA-Sequencer, then it is very common that either Seurat (Hao et al., 2021) or Scanpy (Wolf et al., 2018) packages be used. The quality matric used for preprocessing is based on the method described in the paper (Ilicic et al., 2016).

4.4 Pipeline to identify differentially expressed genes.

The initial procedure in the proposed pipeline is the almost same procedure explained in (Bica et al., 2020). The main difference comes in the selection of top genes from weight distribution. The procedure will be discussed in detail in the following sections.

Hyperparameter setting for variation autoencoder:

Variation autoencoder is built using an encoder and decoder neural network and its architecture is symmetrical in shape. The encoding layer has 512 and 256 neurons in the first and second hidden layers respectively whereas in the decoding layer it is 256 and 512. The summary of for VAE model and its hyperparameter is given in figure 10.

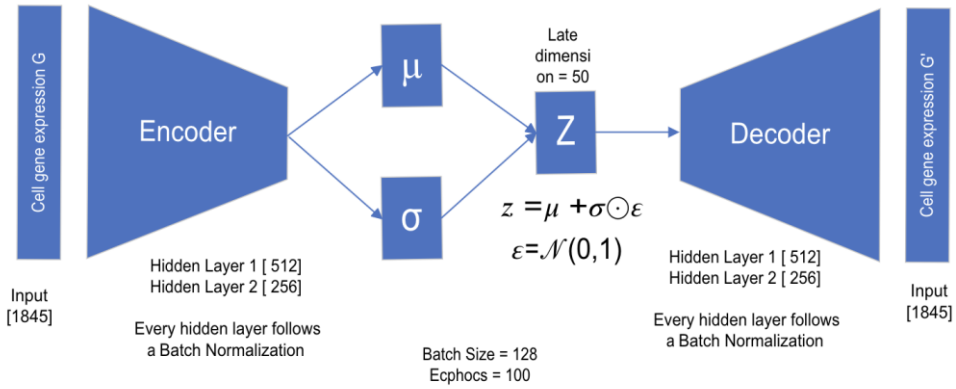


Figure 7: VAE model summary and hyperparameter settings

Dimensionality reduction and clustering

After the training phase gene expression is passed through DiffVAE(Bica et al., 2020) to encode the gene expression into its latent space representation of size 50. This latent representation is gone further through dimensionality reduction techniques using the tSNE method into size 2. Clustering unsupervised method KMEANS algorithm is run over the 2-dimensional space and labeled each cluster by their cluster group as shown in figure 8.

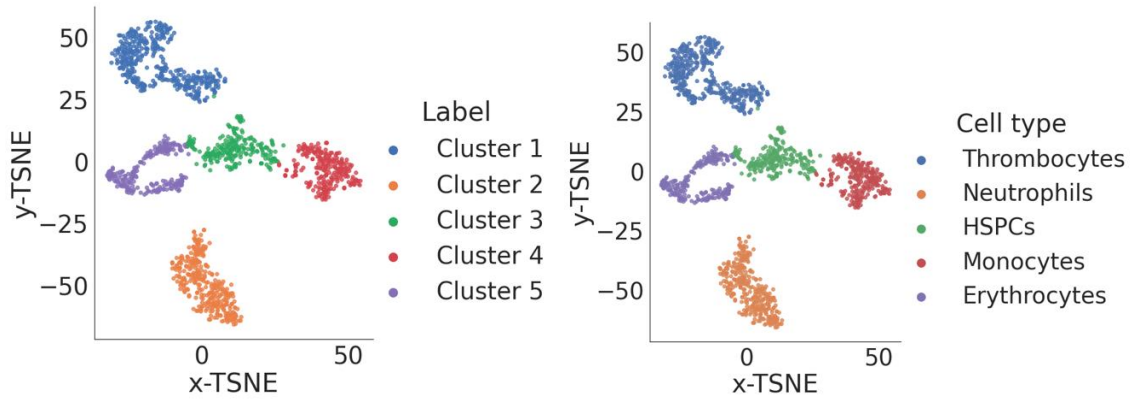


Figure 8: (a) Latent Representation using tSNE (b) label the cluster by their cell type

It may lead to confusion because a latent dimension size for a variational encoder is being set at 50 instead of 2 to avoid using tSNE in the pipeline. Higher dimension size will help in capturing the hidden biological feature in gene expression. Later in the section will discuss in detail how this latent dimension size is useful.

Cell distribution on latent representation.

Measure the distribution of cells on every latent dimension. If we have N number of cells, then every latent dimension can have an array of values with size N.

$$z^k = \left(x_1^k, x_2^k, x_2^k, x_3^k \cdots x_{N-1}^k, x_N^k\right)$$

Let μ^k is mean and σ^k is the variance of the distribution of a latent variable z^k. Filter all the cells with a value greater than $|\mu^k \pm -\sigma^k|$ from z^k (figure 9).

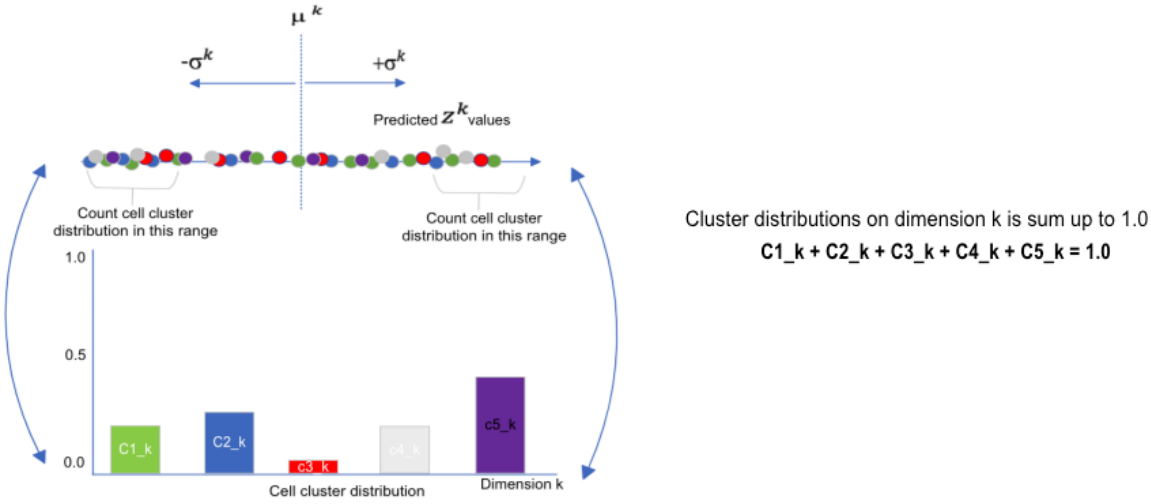


Figure 9: Cluster distribution on latent dimension k

Now we have a certain distribution of cells in each dimension z^k by their cluster types as given the figure 9**.** Identify the highly influencing latent dimension for each cluster. From figure 10, it has been observed that cluster 2, latent dimensions 2, 11, 38, 33, and 40 are the top 5 influential latent dimensions. These dimensions will be responsible for capturing hidden gene expression patterns.

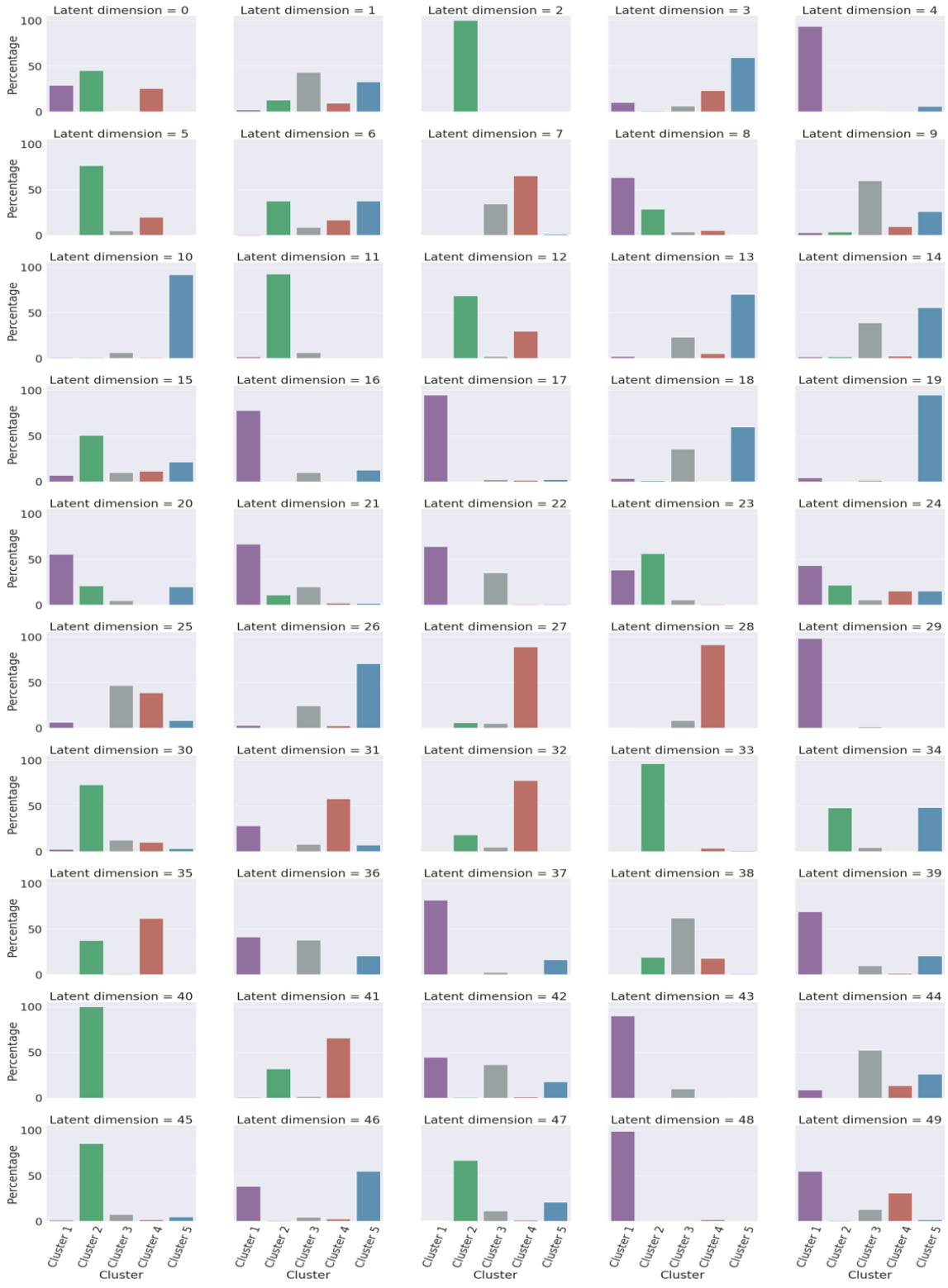


Figure 10: Cluster distributions on the latent dimension

Merging weights in the decoding layer to single layer weighted matrix

Weights associated with each layer in the decoding layers are merged into single weighted matrix that creates a direct weighted relationship between the latent dimension and output gene expression

$$W^0 \in R^{mXn_1},\ W^1 \in R^{n_1Xn_2},\ W^1 \in R^{n_2X\,n}$$
$$\boldsymbol{W} = W^0 \cdot W^1 \cdot W^2$$

From figure 11, $\boldsymbol{W_{ij}}$ indicates the weight of the connection between latent dimension-i and gen-j.

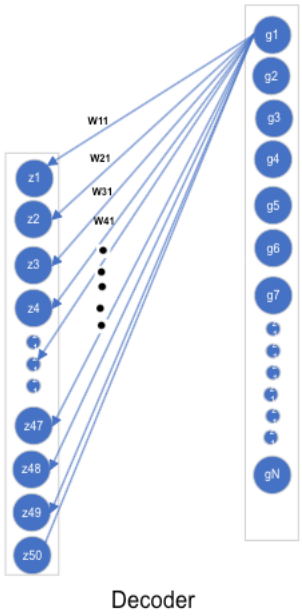
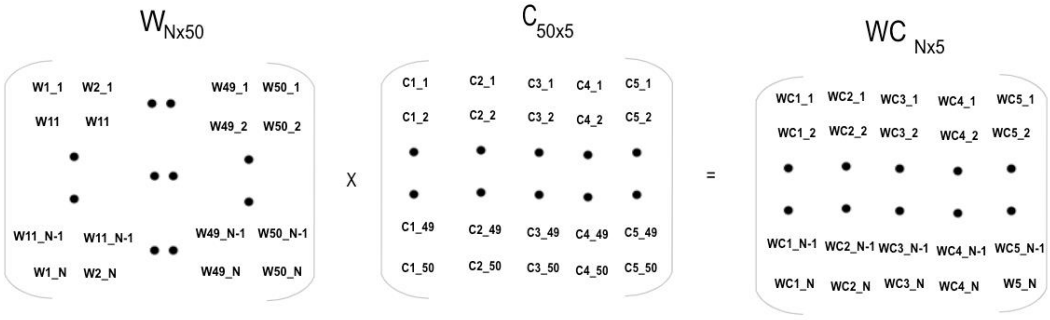


Figure 11: Decoding layer after weight merge

Finding the top-ranked genes that highly differentiate a cluster from other clusters

Let Matrix WC be the weighted sum of weight matrix W and matrix C (distribution percentage of clusters on latent dimensions), cluster weight can be computed as shown in equation 1.



Equation 1: Cluster weight computation

$$WC_{Nx50} = W_{Nx50} \times C_{50x5}$$

Gene-i will be assigned to the cluster with the highest weighted sum.

$$WC_j = \text{MAX}\ (WC_{1j}, WC_{2j}, WC_{3j}, WC_{4j}, WC_{5j})$$

After this stage, all genes will be assigned to a cluster. Genes in the cluster are sorted based on the weight of genes computed. Genes in each of the clusters are ranked by their associated weights WC_j.

4.5 **Analysis.**

In table 3, genes are ordered by their ranks from high to low. Associated logFC values are from the paper (Athanasiadis et al., 2017) and are also shown in the table for validation. But logFC entries were not ordered by decreasing value. Later in this section using visual methods, we can see proposed method is providing a better ordering of the genes than logFC method.

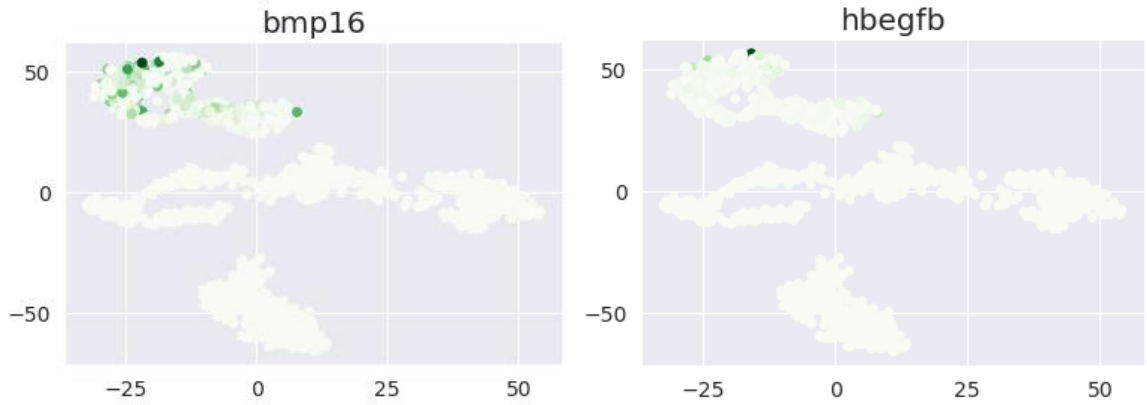
Table 3: Highly computed genes ordered by their ranking,

logfc is from (Athanasiadis et al., 2017)

Cluster1 - (Thrombocytes)	logFC	Cluster2(Neutrophils)	logFC	Cluster3(HSPC)	logFC	Clutser4(Monocytes)	logFC	Cluster5(Erythrocytes)	logFC
fn1b	9.159	lyz	11.961	npm1a	2.98	s100a10b	5.535	si:xx-by187g17.1	7.537
ctgfa	10.249	npsn	10.934	si:ch211-161c3.6	3.576	lgals2a	7.993	ba1l	7.630
gp1bb	5.503	mpx	11.123	cad	3.334	c1qb	8.238	hbaa1	7.480
itga2b	9.2112	ponzr6	8.369	CABZ01070258.1	2.829	c1qc	12.155	ba1	7.447
tuba8l3	11.987	lect2l	9.172	Pcna	1.693	hp	5.502	zgc:92880	7.456
bmp16	9.078	illr4	8.847	Mych	2.123	**si:ch211-165i18.2**		slc4a1a	5.840
rac3a	8.052	cpa5	8.382	eif4a1a	1.780	c1qa	10.325	si:ch211-5k11.6	7.061
tspan7	9.549	cfl1l	8.928	**nanos1**		cfp	4.819	si:ch211-103n10.5	6.832
si:ch211-	7.411	sult2st1	6.769	**fgfrl1a**		**lgals3bpb**		si:dkey-25o16.2	5.302

195b11.3									
si:ch211-214p13.9	7.395	BX908782.3	7.088	**prmt3**		KRT18	5.904	alas2	5.634
thbs1b	7.730	ctssa	9.087	**ascc1**		si:dkey-5n18.1	8.185	cahz	6.374
apln	8.968	thy1	5.900	rhebl1	1.648	slc3a2a	7.328	aqp1a.1	6.148
ADM	9.070	si:ch1073-429i10.1	8.216	**SSBP4**		tspan36	4.051	**dmtn**	4.594
sele	6.216	alox5ap	5.419	**ndst3**		FCER2	5.515	si:ch211-207c6.2	6.874
mpl	5.381	gapdh	8.208	tnrc18	2.898	grna	6.908	igfbp1a	3.874
si:dkeyp-116a7.2	4.054	abcb9	7.555	**si:dkey-257i7.5**		**irf8**		nt5c2l1	4.464
blf	3.727	PLPP1	5.381	**si:ch211-250k18.7**		**si:ch211-283g2.2**		si:ch211-250g4.3	5.245

From the paper (Athanasiadis et al., 2017) logFC value for gene (hbegfb) is 9.881 which belongs to cluster 1(Thrombocytes) and comes within the top 5. But according to our ranking, it comes at 47th position. Our ranking is reasonably working well as you see in figure 14. Based on the ranking, using the tool I developed gene (bmp16) comes in 5th position. It is clear from figure 11 that the gene (bmp16) is more highly expressed than the gene (hbegfb). Their difference in the ranking also shows how significantly they differ. You can see in figure 11, that logFC is not able to capture some genes that differentiate from another cluster. Examples of such genes are (si:ch211-165i18.2) and (**l**gals3bpb) which are belongs to cluster 4.



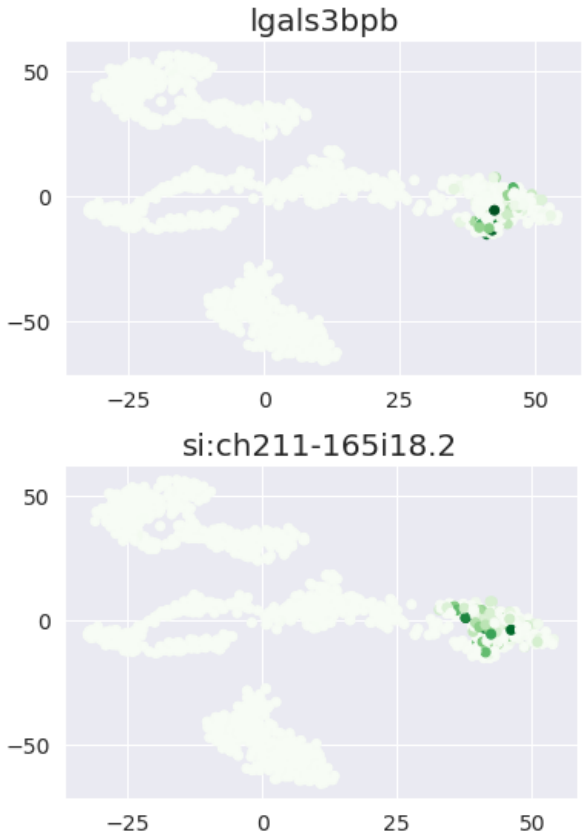


Figure 12: Gene Expression of gene

bmp16, hbegfb, lgals3bpb, si:ch211-165il8.2

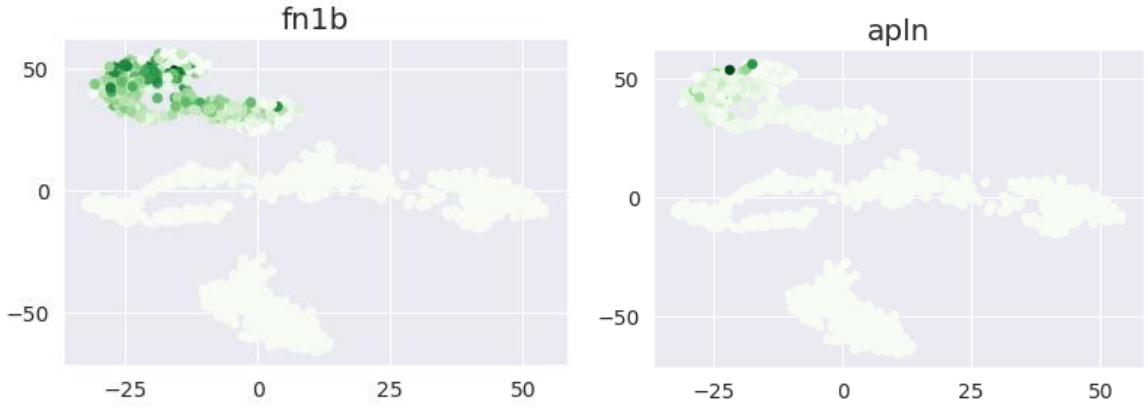
As stated, before our reference data is from (Bica et al., 2020) and I did a performance comparison with top genes published in that paper. Table 4 lists the top genes for each cluster published by paper (Bica et al., 2020) which differentiate a cluster from another cluster group. I give a ranking for those genes based on our tool. In every cluster, it is observed missing values in the ordered gene ranking.

Table 4: Highly expresses genes found by (Bica et al., 2020) and provided the ranking

Cluster1-gene	Rank	Cluster2-gene	Rank	Cluster3-gene	Rank	Cluster4-gene	Rank	Cluster5-gene	Rank
fn1b	0	lyz	0	npm1a	0	s100a10b	0	si:xx-by187g17.1	0
ctgfa	1	npsn	1	si:ch211-161c3.6	1	lgals2a	1	ba1l	1
itga2b	3	ponzr6	3	cad	2	c1qc	3	hbaa1	2

bmp16	5	illr4	5	pcna	4	hp	4	ba1	3
rac3a	6	cpa5	6	nanos1	7	si:ch211-165i18.2	5	zgc:92880	4
tspan7	7	cfl1l	7	fgfrl1a	8	lgals3bpb	8	si:ch211-5k11.6	6
si:ch211-195b11.3	8	ctssa	10	prmt3	9	FCER2	13	si:ch211-103n10.5	7
thbs1b	1	si:ch1073-429i10.1	12			si:ch211-283g2.2	16	si:dkey-25o16.2	8
apln	11	hsd3b7	17			cd74a	10	alas2	9
blf	16	scpp8	18			marco	20	dmtn	12
pmp22b	21	mmp13a	20			zgc:136870	53	si:ch211-207c6.2	13
rhag	31	cfd	47			zgc:136870	53	igfbp1a	14
bmp6	48	ANPEP	53					rgcc	18
tnr	50	mmp9	54					mibp	19
fhl2a	66	Aspm	3118					si:ch211-197g15.10	23
		kif11	436						

In the above table 4 gene (fn1b) and (rhag) which are belongs to cluster 1. The table is listing top 15 genes. But the difference in the ranking as per my tool is 31. In my top listed genes in table 3, the 11th ranked gene is (apln). The comparison of genes (fn1b, rhag and apln) expressions are given the figure 12. This shows that the modified algorithm works very well than the original algorithm explained in (Bica et al., 2020)



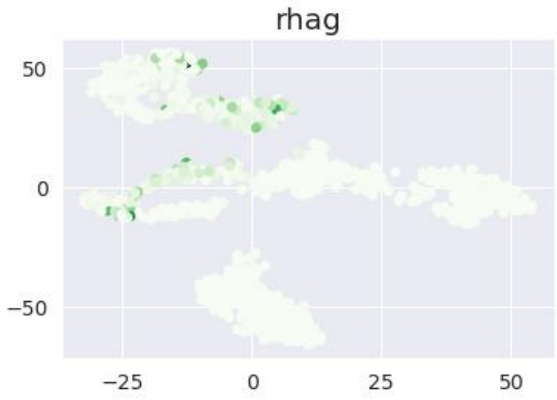


Figure 13: Gene Expression of fn1b, apln and rhag

Visual comparison of top listed genes listed with genes listed in (Bica et al., 2020)

Here I am doing a performance comparison using the visual method to show how my top listed genes were expressed and ranked based on their significance in cluster make. By giving ranking to all genes that were published as top listed genes in the paper Bica et al., 2020, we can measure the quality level of every gene in the cluster make. The following section will go to each cluster for performance comparison.

Cluster 1:

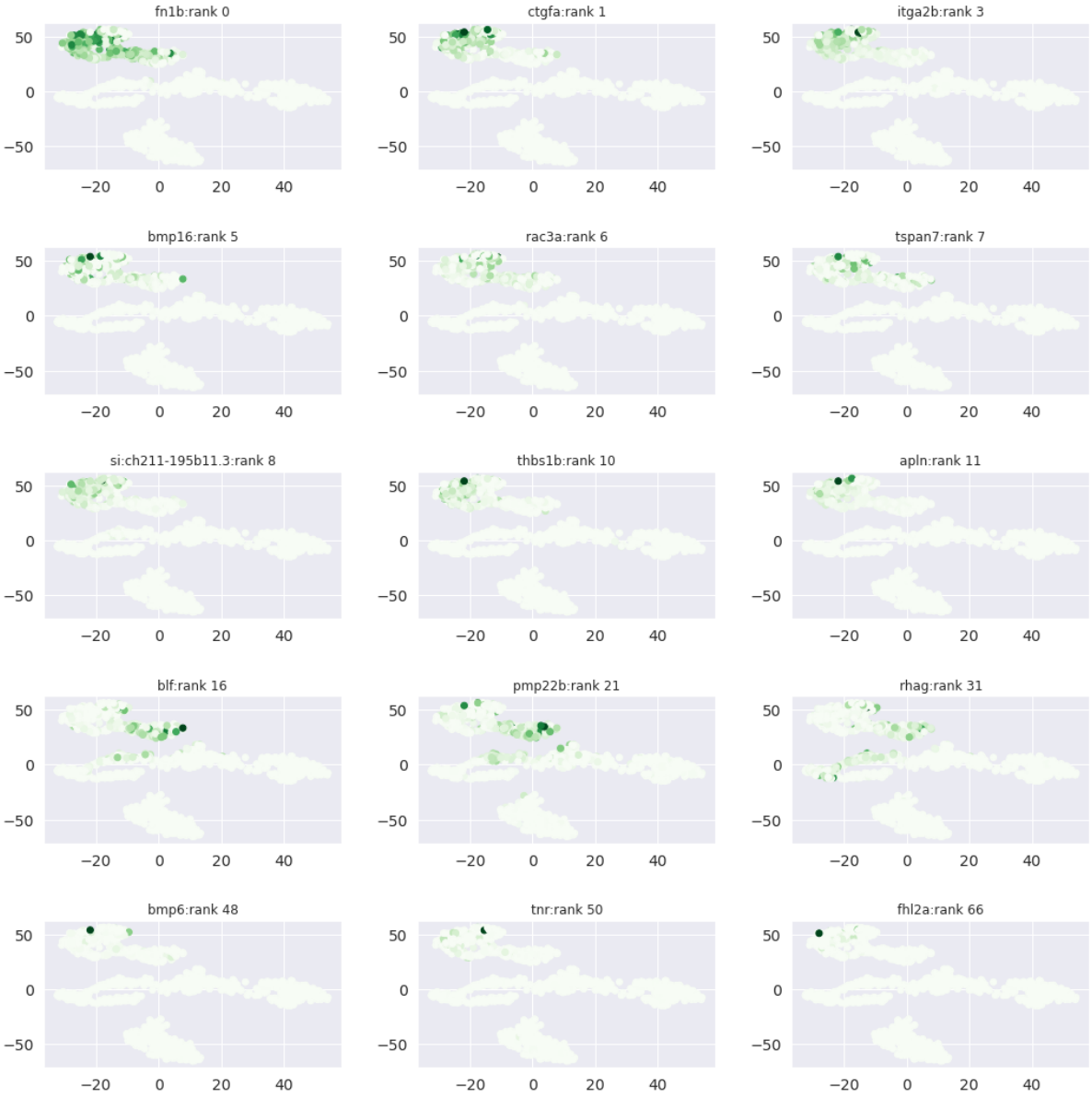


Figure 14: Cluster 1: Gene Expression of top listed genes by (Bica et al., 2020)

Figure 15: Cluster 1: Gene expression of a new set of top listed genes discovered

Cluster 2:

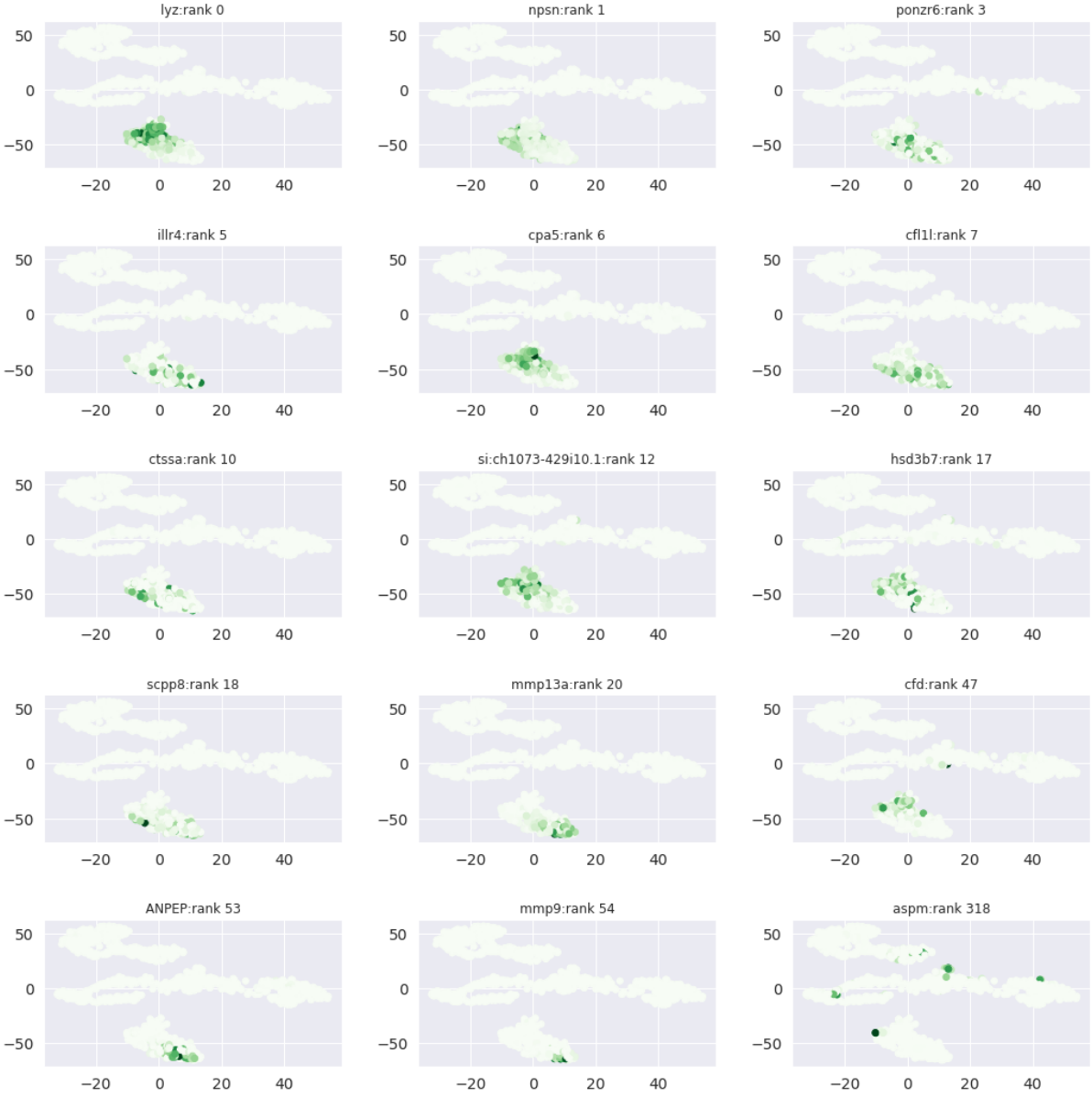


Figure 16: Cluster 2: Gene Expression of top listed genes by (Bica et al., 2020)

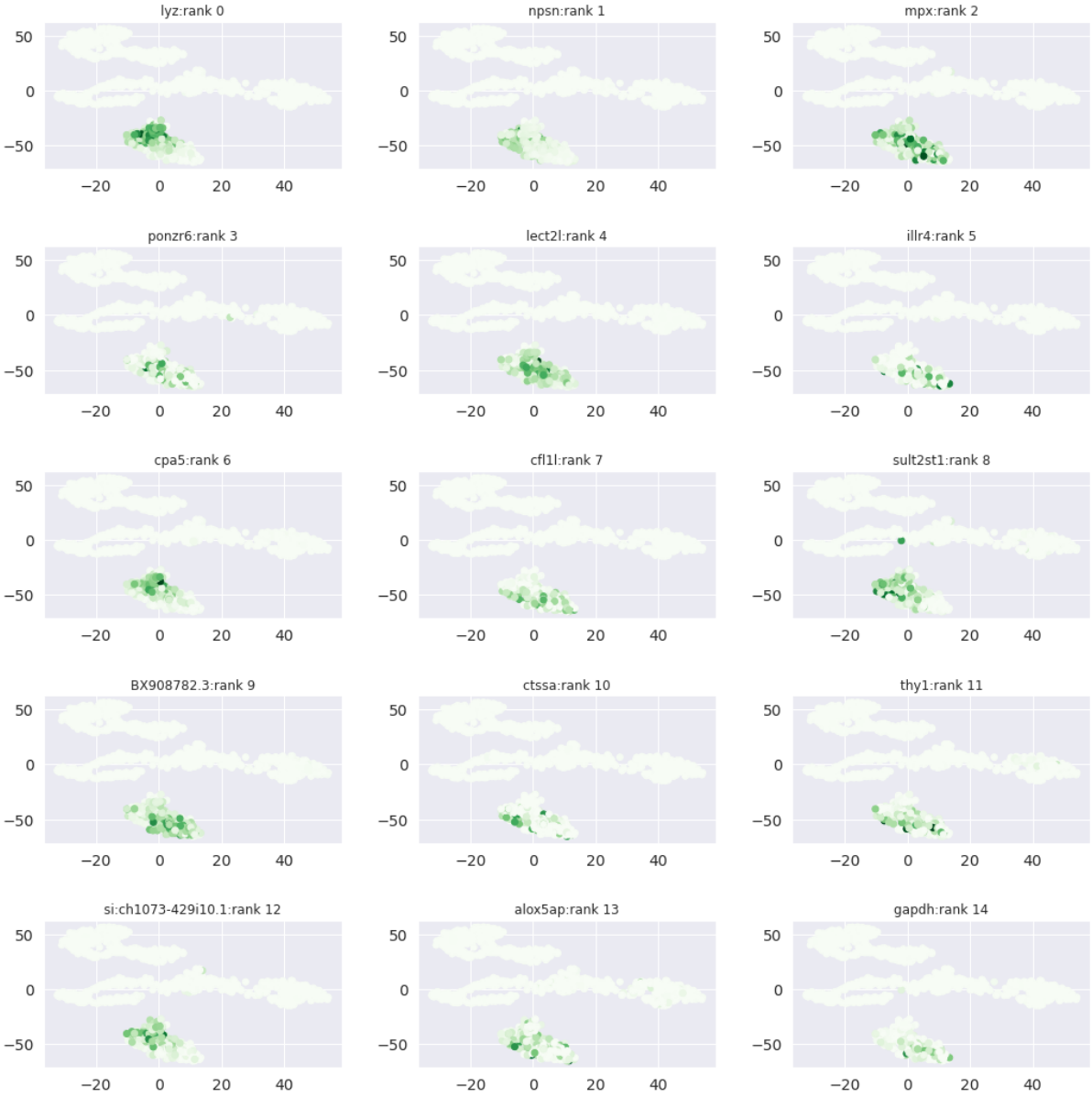


Figure 17: Cluster 2: Gene expression of new set top listed gene discovered

CHAPTER 5: CONCLUSION AND FUTURE WORK

As stated, earlier BiDiffVAE is an extension of DiffVAE to find the marker genes. In DiffVAE it gets its maker gene, by merging all top genes associated with latent dimensions. Merging of genes will be done on the same type of latent dimension. Hence one gene can be assigned to multiple cluster groups. This way it is looking at genes from each latent dimension space. But in BiDiffVAE one gene is assigned to one cluster group as the gene is looking at all latent space together to compute cluster weightage to assign respective cluster type. This computed cluster weightage is used for ranking genes in clusters. From our result, it has been observed genes with higher ranks gene expressiveness in one cluster is high and in the other cluster is very low. This expressiveness will decrease for low-ranking genes

Here are my research questions and their answers.

- How to include the cluster wights in finding differentially expressed genes?
  BiDiffVAE followed the basic computational model from DiffVAE. Weights in the decoding layer merged first and computed cluster distribution in every latent dimension. But in BiDiffVAE genes' weights were multiplied by the corresponding cluster distribution.
- How does including the cluster distribution improve the performance?
  In BiDiffVAE all weights associated with a gene are added together. But this tool is expecting a different total sum for different cluster types. To achieve this all weights are multiplied by the corresponding cluster distribution. BiDiffVAE this way accounts for all latent dimensions together while assigning cluster type and outperforms in finding high quality differentially expressed genes.

This opens up further research, since there is no gold model available to compare the results with the tool I proposed, it would be nice to have simulated/synthetic data that can be run against other popular statistical tools. It would be nice to try techniques used exclusively in the BiDiffVAE to those models covered in the related work section. Since this tool discovered new sets of genes it will open a new set of research questions for the researcher for the downstream analysis.

REFERENCES

Arisdakessian, C., Poirion, O., Yunits, B., Zhu, X., Garmire, L.X., 2019. Genome Biology 20.

Athanasiadis, E.I., Botthof, J.G., Andres, H., Ferreira, L., Lio, P., Cvejic, A., 2017. Nature Communications 2017 8:1 8, 1–11.

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I.W.H., Ng, L.G., Ginhoux, F., Newell, E.W., 2018. Nature Biotechnology 2018 37:1 37, 38–44.

Bianconi, E., Piovesan, A., Facchin, F., Beraudi, A., Casadei, R., Frabetti, F., Vitale, L., Pelleri, M.C., Tassani, S., Piva, F., Perez-Amodio, S., Strippoli, P., Canaider, S., 2013. Annals of Human Biology 40, 463–471.

Bica, I., Andrés-Terré, H., Cvejic, A., Liò, P., 2020. Scientific Reports 2020 10:1 10, 1–13.

Bondoc, A., Glaser, K., Jin, K., Lake, C., Cairo, S., Geller, J., Tiao, G., Aronow, B., 2021. Communications Biology 2021 4:1 4, 1–14.

Brbić, M., Zitnik, M., Wang, S., Pisco, A.O., Altman, R.B., Darmanis, S., Leskovec, J., 2020. Nature Methods 2020 17:12 17, 1200–1206.

Chen, G., Ning, B., Shi, T., 2019. Frontiers in Genetics 10, 317.

Deng, Y., Bao, F., Dai, Q., Wu, L.F., Altschuler, S.J., 2019. Nature Methods 2019 16:4 16, 311–314.

Ding, J., Condon, A., Shah, S.P., 2018. Nature Communications 2018 9:1 9, 1–13.

Eraslan, G., Simon, L.M., Mircea, M., Mueller, N.S., Theis, F.J., 2019. Nature Communications 2019 10:1 10, 1–14.

Fang, R., Preissl, S., Li, Y., Hou, X., Lucero, J., Wang, X., Motamedi, A., Shiau, A.K., Zhou, X., Xie, F., Mukamel, E.A., Zhang, K., Zhang, Y., Behrens, M.M., Ecker, J.R., Ren, B., 2021. Nature Communications 2021 12:1 12, 1–15.

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A.K., Slichter, C.K., Miller, H.W., McElrath, M.J., Prlic, M., Linsley, P.S., Gottardo, R., 2015. Genome Biology 16.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E.P., Jain, J., Srivastava, A., Stuart, T., Fleming, L.M., Yeung, B., Rogers, A.J., McElrath, J.M., Blish, C.A., Gottardo, R., Smibert, P., Satija, R., 2021. Cell 184, 3573-3587.e29.

Hwang, B., Lee, J.H., Bang, D., 2018. Experimental and Molecular Medicine 50, 1–14.

Ilicic, T., Kim, J.K., Kolodziejczyk, A.A., Bagger, F.O., McCarthy, D.J., Marioni, J.C., Teichmann, S.A., 2016. Genome Biology 17.

Kamath, T., Abdulraouf, A., Burris, S.J., Langlieb, J., Gazestani, V., Nadaf, N.M., Balderrama, K., Vanderburg, C., Macosko, E.Z., 2022. Nature Neuroscience 2022 25:5 25, 588–595.

Liu, B., Li, C., Li, Z., Wang, D., Ren, X., Zhang, Z., 2020. Nature Communications 2020 11:1 11, 1–13.

Lotfollahi, M., Naghipourfar, M., Luecken, M.D., Khajavi, M., Büttner, M., Wagenstetter, M., Avsec, Ž., Gayoso, A., Yosef, N., Interlandi, M., Rybakov, S., Misharin, A. v., Theis, F.J., 2021. Nature Biotechnology 2021 40:1 40, 121–130.

Lotfollahi, M., Naghipourfar, M., Theis, F.J., Alexander Wolf, F., 2020. Bioinformatics 36, I610–I617.

Lotfollahi, M., Wolf, F.A., Theis, F.J., 2019. Nature Methods 16, 715–721.

Love, M.I., Huber, W., Anders, S., 2014. Genome Biology 15, 1–21.

Luecken, M.D., Theis, F.J., 2019. Molecular Systems Biology 15, e8746.

Lukassen, S., Ten, F.W., Adam, L., Eils, R., Conrad, C., 2020. Nature Machine Intelligence 2020 2:12 2, 800–809.
van der Maaten, L., Hinton, G., 2008. Journal of Machine Learning Research 9, 2579–2605.
McDavid, A., Finak, G., Chattopadyay, P.K., Dominguez, M., Lamoreaux, L., Ma, S.S., Roederer, M., Gottardo, R., 2013. Bioinformatics 29, 461–467.
Prokaryotes vs Eukaryotes [WWW Document], n.d. URL https://www.technologynetworks.com/cell-science/articles/prokaryotes-vs-eukaryotes-what-are-the-key-differences-336095 (accessed 5.4.22).
Seninge, L., Anastopoulos, I., Ding, H., Stuart, J., 2021. Nature Communications 12, 1–9.
Shu, H., Zhou, J., Lian, Q., Li, H., Zhao, D., Zeng, J., Ma, J., 2021. Nature Computational Science 2021 1:7 1, 491–501.
Sidhom, J.W., Larman, H.B., Pardoll, D.M., Baras, A.S., 2021. Nature Communications 2021 12:1 12, 1–12.
Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y., Stoeckius, M., Smibert, P., Satija, R., 2019. Cell 177, 1888-1902.e21.
Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P., 2005a. Proc Natl Acad Sci U S A 102, 15545–15550.
Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P., 2005b. Proc Natl Acad Sci U S A 102, 15545–15550.
Svensson, V., Gayoso, A., Yosef, N., Pachter, L., 2020. Bioinformatics 36, 3418–3421.
Svensson, V., Vento-Tormo, R., Teichmann, S.A., 2018. Nature Protocols 2018 13:4 13, 599–604.
The Information in DNA Determines Cellular Function via Translation | Learn Science at Scitable [WWW Document], n.d. URL https://www.nature.com/scitable/topicpage/the-information-in-dna-determines-cellular-function-6523228/ (accessed 5.4.22).
Tian, T., Wan, J., Song, Q., Wei, Z., 2019. Nature Machine Intelligence 2019 1:4 1, 191–198.
What are proteins and what do they do?: MedlinePlus Genetics [WWW Document], n.d. URL https://medlineplus.gov/genetics/understanding/howgeneswork/protein/ (accessed 5.9.22).
Wolf, F.A., Angerer, P., Theis, F.J., 2018. Genome Biology 19, 1–5.
Xu, C., Lopez, R., Mehlman, E., Regier, J., Jordan, M.I., Yosef, N., 2021. Molecular Systems Biology 17, 9620.
Zhao, Y., Cai, H., Zhang, Z., Tang, J., Li, Y., 2021. Nature Communications 2021 12:1 12, 1–15.

APPENDICES

Cluster 3:

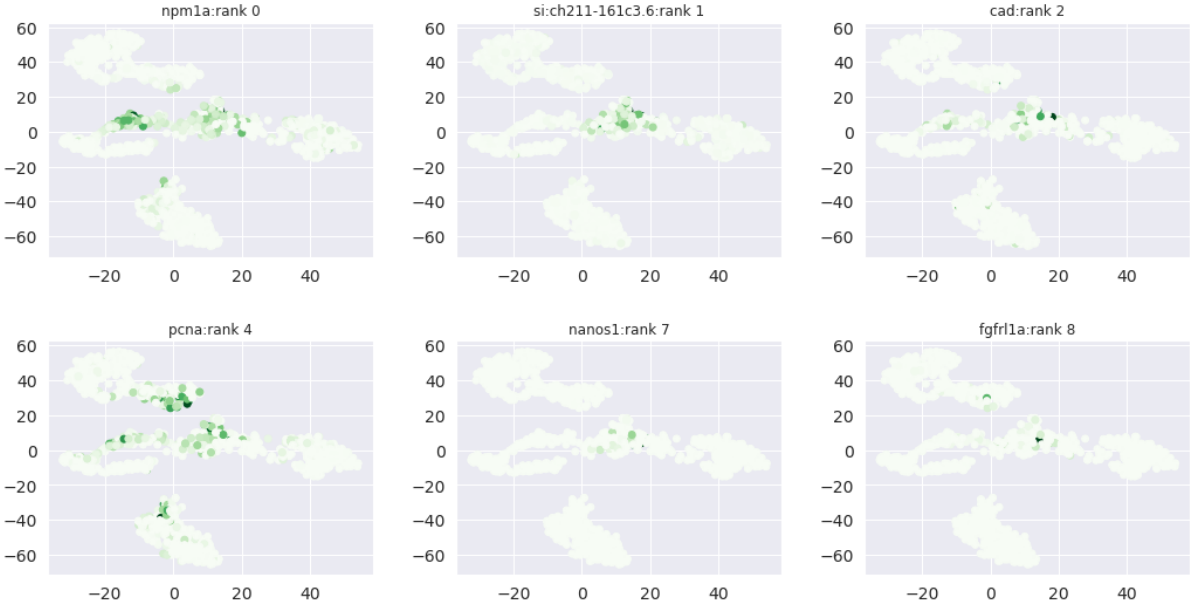


Figure 18: Cluster 3: Gene Expression of top listed genes by (Bica et al., 2020)

Figure 19: Cluster 3: Gene expression of new set top listed gene discovered

Cluster 4:

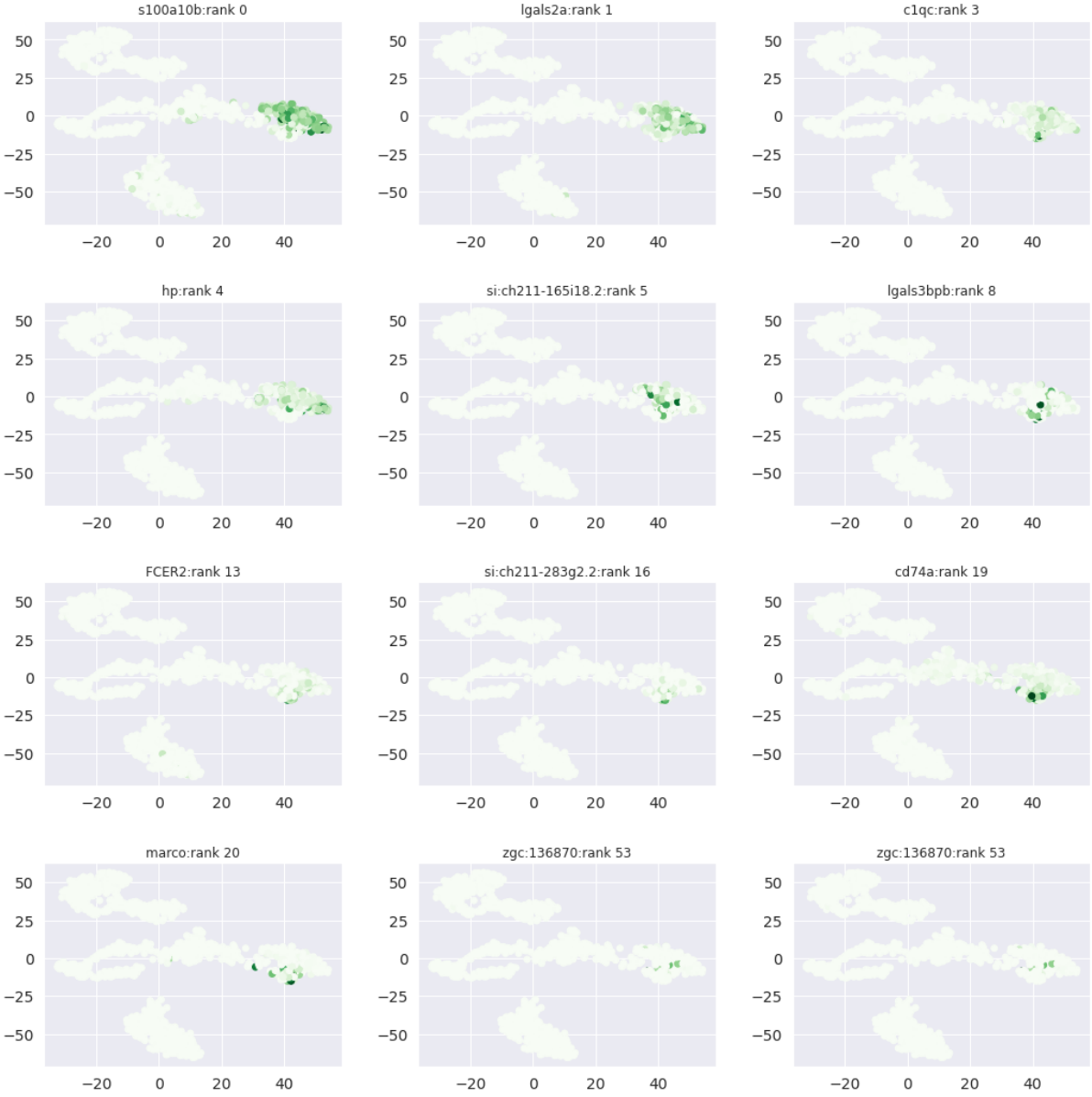


Figure 20: Cluster 4: Gene Expression of top listed genes by (Bica et al., 2020)

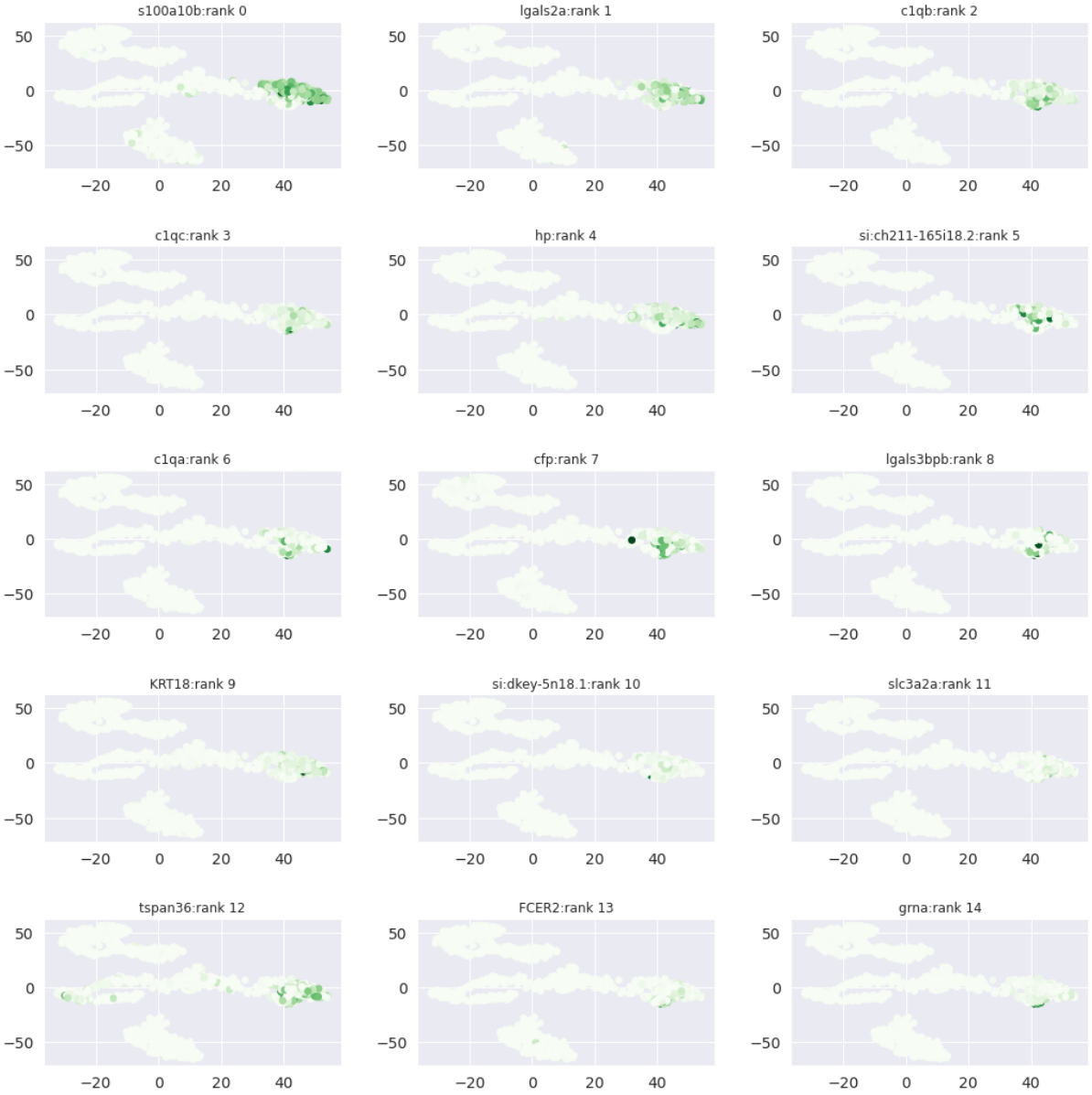


Figure 21: Cluster 4: Gene expression of new set top listed gene discovered

Cluster 5:

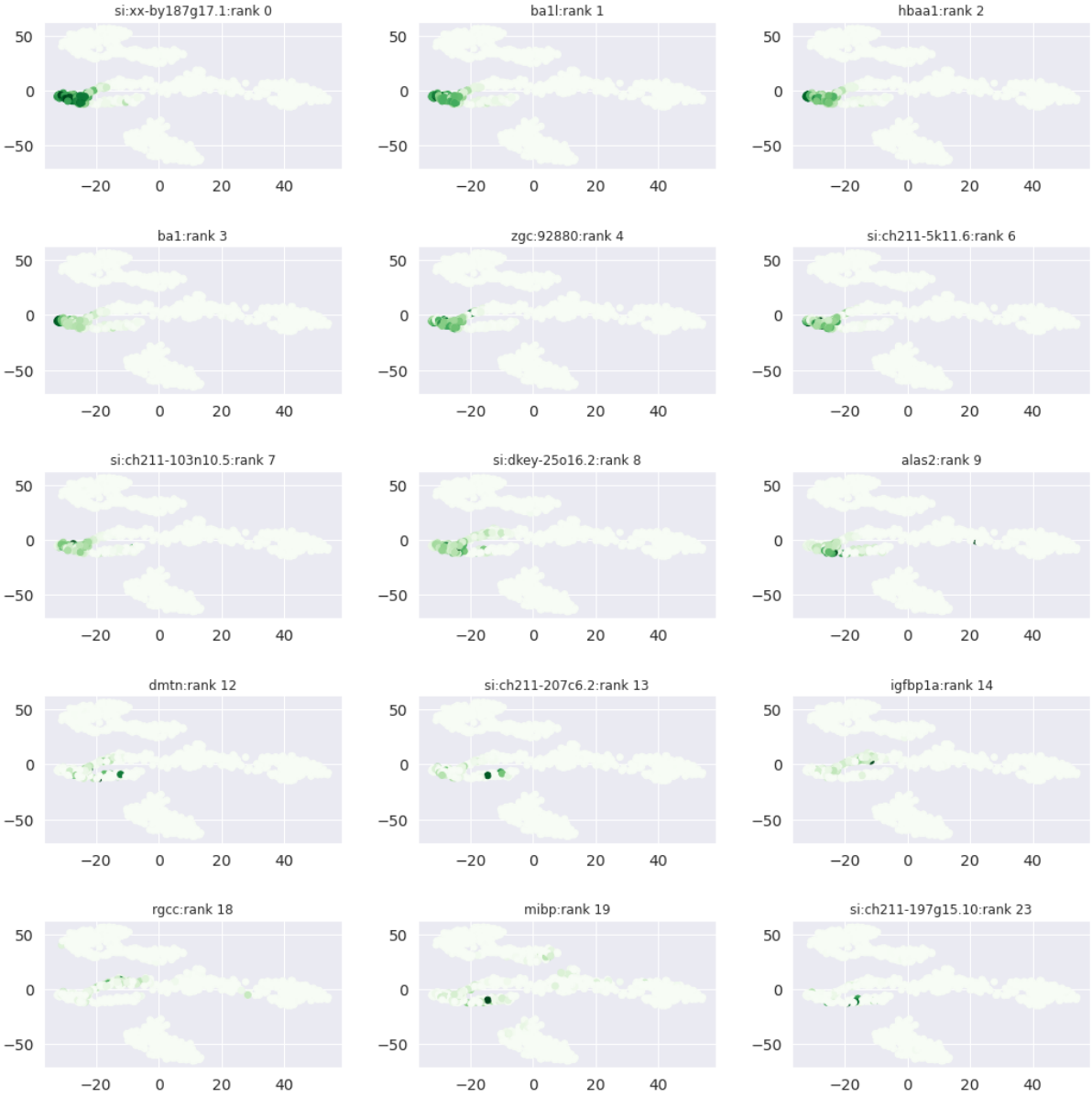


Figure 22: Cluster 5: Gene Expression of top listed genes by (Bica et al., 2020)

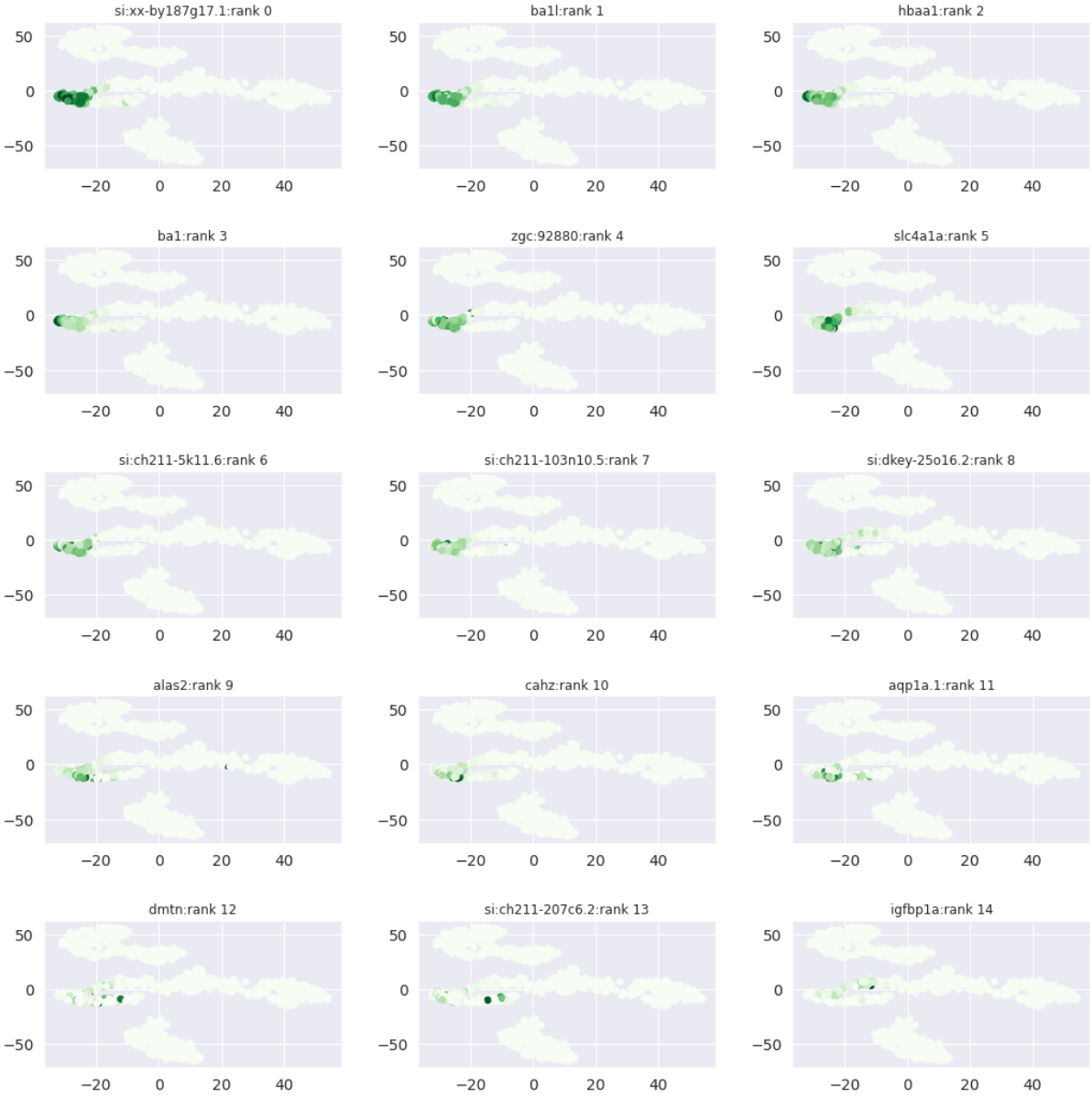


Figure 23: Cluster 5: Gene expression of new set top listed gene discovered