University of Windsor

Scholarship at UWindsor

6-18-2021

Identification of Cell Types in scRNA-seq Data via Enhanced Local Embedding and Clustering

Saiteja Danda
University of Windsor

Follow this and additional works at: https://scholar.uwindsor.ca/etd

Recommended Citation

Danda, Saiteja, "Identification of Cell Types in scRNA-seq Data via Enhanced Local Embedding and Clustering" (2021). *Electronic Theses and Dissertations*. 8592.
https://scholar.uwindsor.ca/etd/8592

This online database contains the full-text of PhD dissertations and Masters' theses of University of Windsor students from 1954 forward. These documents are made available for personal study and research purposes only, in accordance with the Canadian Copyright Act and the Creative Commons license—CC BY-NC-ND (Attribution, Non-Commercial, No Derivative Works). Under this license, works must always be attributed to the copyright holder (original author), cannot be used for any commercial purposes, and may not be altered. Any other use would require the permission of the copyright holder. Students may inquire about withdrawing their dissertation and/or thesis from this database. For additional inquiries, please contact the repository administrator via email (scholarship@uwindsor.ca) or by telephone at 519-253-3000ext. 3208.

Identification of Cell-types in scRNA-seq Data via Enhanced Local Embedding and Clustering

By

Saiteja Danda

A Thesis
Submitted to the Faculty of Graduate Studies
through the School of Computer Science
in Partial Fulfillment of the Requirements for
the Degree of Master of Science
at the University of Windsor

Windsor, Ontario, Canada

2021

©2021 Saiteja Danda

Identification of Cell-types in scRNA-seq Data via Enhanced Local Embedding and Clustering

by

Saiteja Danda

APPROVED BY:

P. Karpowicz
Department of Biomedical Sciences

A. Biniaz
School of Computer Science

L. Rueda, Advisor
School of Computer Science

April 22, 2021

DECLARATION OF CO-AUTHORSHIP / PREVIOUS PUBLICATION

I. Co-Authorship

I hereby declare that this thesis incorporates material that is result of joint research, as follows:

Chapter 2 of the thesis was co-authored with Akram Vasighizaker and Luis Rueda. L. Rueda contributed with initial thoughts about this research area and the main ideas, and assisted in elaborating on the new novel approach implemented in this work. All the authors contributed in finalizing the ideas and follow-up discussions. A. Vasighizaker helped in the data collection step and biological assessment. S. Danda implemented the method, data pre-processing, experimental design, and the data analysis. S. Danda, A. Vasighizaker, and L. Rueda wrote the contents of the chapter. L. Rueda supervised the project.

I am aware of the University of Windsor Senate Policy on Authorship and I certify that I have properly acknowledged the contribution of other researchers to my thesis, and have obtained written permission from each of the co-author(s) to include the above material(s) in my thesis. I certify that, with the above qualification, this thesis, and the research to which it refers, is the product of my own work.

II. Previous Publication

This thesis includes two original papers, one of which has been published, while the other is being submitted to a scientific journal, as follows:

Thesis chapter	**Publication title**	**Publication Status**
Chapter 2 (primary version)	S. Danda, A. Vasighizaker and L. Rueda, ”Unsupervised Identification of SARS-CoV-2 Target Cell Groups via Nonlinear Dimensionality Reduction on Single-cell RNA-Seq Data,” 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Seoul, Korea (South), 2020, pp. 2737-2744, doi: 10.1109/BIBM49941.2020.9313378.	Published
Chapter 2 (Extended)	A. Vasighizaker, S. Danda and L. Rueda, Discovering Cell Types Using Manifold Learning and Enhanced Visualization of Single-cell RNA-Seq Data	To be submitted

I certify that I have obtained a written permission from the copyright owner(s) to include the above published material(s) in my thesis. I certify that the above material describes work completed during my registration as a graduate student at the University of Windsor.

III. General

I declare that, to the best of my knowledge, my thesis does not infringe upon anyone’s copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my thesis and have included copies of such copyright clearances to my appendix.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution. I understand that my thesis may be made electronically available to the public.

ABSTRACT

Identifying specific cell types is a significant step for studying diseases and potentially leading to better diagnosis, drug discovery, and prognosis. High-throughput single-cell RNA-Seq (scRNA-seq) technologies have advanced in recent years, enabling researchers to investigate cells individually and understand their biological mechanisms. Computational techniques such as clustering, which are categorized in the form of unsupervised learning methods, are the most suitable approach in scRNA-seq data analysis when the cell types have not been characterized. These techniques can be used to identify a group of genes that belong to a specific cell type based on their similar gene expression patterns. However, due to the sparsity and high-dimensional nature of scRNA-seq data, classical clustering methods are not efficient. Therefore, the use of non-linear dimensionality reduction techniques to improve clustering results is crucial. We introduce a pipeline to identify representative clusters of different cell types by combining non-linear dimensionality reduction techniques such as modified locally linear embedding (MLLE) and clustering algorithms. We assess the impact of different dimensionality reduction techniques combined with the clustering of thirteen publicly available scRNA-seq datasets of different tissues, sizes, and technologies. We evaluate the intra- and inter-cluster performance based on the Silhouette score before performing a biological assessment. We further performed gene enrichment analysis across biological databases to evaluate the proposed method's performance. As such, our results show that MLLE combined with independent component analysis yields overall the best performance relative to the existing unsupervised methods across different experiments.

Keywords: non-linear dimensionality reduction, clustering, single-cell RNA sequencing, cell type identification, unsupervised learning.

DEDICATION

I would like to dedicate this thesis to my parents for their endless love, support and encouragement.

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Dr. Luis Rueda, for his exceptional guidance, patience, and constant support. This thesis would have not been possible without his ideas and suggestions. I have greatly benefited from his immense knowledge. It has been such an honour for me to work under his supervision.

I would also like to thank my external reader Dr. Phillip Karpowicz and internal reader Dr. Ahmad Biniaz for being a part of my thesis committee. Their feedback and comments have helped in further improving of this thesis.

Finally, I would like to thank all my friends and family those who have been a source of constant morale support specially Vinay, who motivated me to choose AI and always backed me, all my lab-mates at the university, my sisters, Aekta, Steffie, Akram, Unnadi Okate Zindagi group, Gajapade, and all my school friends.

TABLE OF CONTENTS

DECLARATION OF CO-AUTHORSHIP / PREVIOUS PUBLICATION III

ABSTRACT V

DEDICATION VI

ACKNOWLEDGEMENTS VII

LIST OF TABLES X

LIST OF FIGURES XI

1 Introduction 1
- 1.1 Introduction to Molecular Biology . . . 1
- 1.2 Central Dogma of Molecular Biology . . . 2
- 1.3 Gene Expression . . . 3
- 1.4 RNA Sequencing . . . 4
- 1.5 Single-cell RNA sequencing (scRNA-seq) . . . 5
- 1.6 Machine Learning . . . 7
  - 1.6.1 Unsupervised Learning . . . 7
  - 1.6.2 Cluster Analysis . . . 8
  - 1.6.3 *k*-Means Algorithm . . . 8
  - 1.6.4 Dimensionality Reduction . . . 9
- 1.7 Motivation . . . 12
- 1.8 Problem Statement . . . 13
- 1.9 Proposed Method . . . 13
  - 1.9.1 Contributions . . . 14
- References . . . 14

2 Manifold Learning and Enhanced Visualization of scRNA-seq data 18
- 2.1 Introduction . . . 18
- 2.2 Materials and Methods . . . 20
- 2.3 Materials and Methods . . . 21
  - 2.3.1 Datasets . . . 21
  - 2.3.2 Data Pre-processing and Quality Control . . . 22
  - 2.3.3 Dimensionality Reduction . . . 25
    - 2.3.3.1 Modified Locally Linear Embedding . . . 27
    - 2.3.3.2 Independent Component Analysis . . . 29
    - 2.3.3.3 Other Dimensionality Reduction Methods . . . 29
  - 2.3.4 Clustering . . . 30
  - 2.3.5 Cluster Annotation . . . 31
  - 2.3.6 Parameter Optimization . . . 31

2.3.7 Performance Evaluation 32
2.4 Results and Discussion 34
2.4.1 Clustering and Cell Type Discovery 34
2.4.2 Biological Assessment 40
References 44

3 Conclusion and Future Work 49
3.1 Conclusion 49
3.1.1 Contributions 49
3.2 Future Work 50
References 51

APPENDIX A Marker Genes 52

VITA AUCTORIS 54

LIST OF TABLES

2.3.1 Datasets used in this work. 22
2.4.1 Parameters used for experiments. These are generated considering both dimensionality reduction and clustering together. 35
2.4.2 Silhoutte scores comparison of proposed method with other dimensionality reduction techniques. 36
2.4.3 Identified cell types for H1299_scRNAseq. 41
2.4.4 Identified cell types for Calu3_scRNAseq. 42
2.4.5 Identified cell types for Baron_human1 dataset. 42

A.0.1 Marker genes found in similar diseases. 52

LIST OF FIGURES

1.1.1 Structure of DNA [**dna˙image**] 1

1.2.1 Central dogma of molecular biology [**central˙image**]. 2

1.4.1 Overview of RNA-seq workflow [**rna˙image**]. 4

1.5.1 Overview of scRNA-seq workflow [**sc˙image**]. 6

1.6.1 Dimensionality reduction from three dimensions to two dimensions using manifold techniques [**dr˙image**]. 11

2.3.1 Block diagram of the proposed approach for discovering cell types in scRNA-seq data. 23

2.3.2 Investigating the distribution of the data to filtered out weakly expressed genes and low-quality cells from dataset; (a) number of expressed genes, (b) total counts per cell, and (c) the percentage of mitochondrial genes for H1299_scRNAseq. 23

2.3.3 Top 20 highly-variable genes before normalization. 24

2.3.4 Top 20 highly-variable genes after normalization. 25

2.3.5 Dispersion of genes before normalization. 26

2.3.6 Dispersion of genes after normalization. 26

2.3.7 Comparison of dispersion of normalized and not normalized genes to extract highly variable genes. 26

2.4.1 Visualization of t-SNE on Muraro dataset 36

2.4.2 Visualization of PCA on Wang dataset 37

2.4.3 Visualization of Laplacian eigenmaps on H1299_scRNAseq; outliers have been removed to enhance visualization. 37

2.4.4 Two-dimensional ICA projection of cells colored by *k*-means clustering applied on high-dimensional original data (H1299_scRNAseq). 38

2.4.5 Two-dimensional ICA projection of cells colored by *k*-means clustering applied to the three-dimensional points output by MLLE on the H1299_scRNAseq dataset. 38

2.4.6 Cluster annotation for H1299_scRNAseq. 39

2.4.7 A set of biological process that are enriched by marker genes in H1299_scRNAseq dataset. The numbers show the clusters and edges shows the link between a cluster and a biological process term. 43

2.4.8 Pathway that are enriched by marker genes in H1299_scRNAseq dataset. The numbers show the clusters and edges shows the link between a cluster and a pathway. Node that is highlighted yellow show the SARS-CoV-2 cell-specific pathway. Most of the other green nodes reveal the shared and cluster-specific functional pathways in the immune system. 43

CHAPTER 1

Introduction

1.1 Introduction to Molecular Biology

Molecular biology, a branch of biology, is the study of the molecular basis of biological activities. A cell is the basic unit of all living organisms. There are two major cell types: eukaryotic cells and prokaryotic cells. Cells with the real nucleus are eukaryotic, and cells with no real nucleus are prokaryotic cells. Hence, eukaryotes and prokaryotes, respectively. Eukaryotic[14] cells consist of many biomolecules such as proteins and nucleic acids. Molecular biologists conduct experiments to find information about the structure, processing, function, regulation, and evolution of biological molecules. Their interactions with one another provide more insights into how life works.

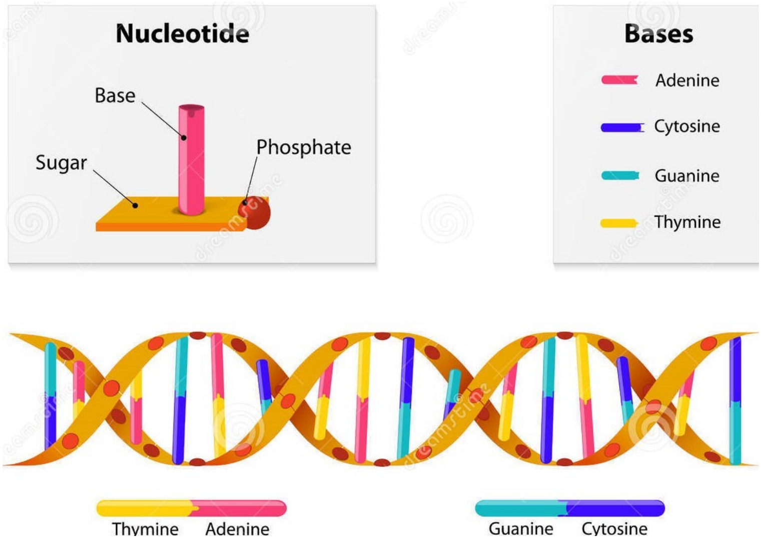

Fig. 1.1.1: Structure of DNA [10]

DNA (Deoxy-Ribo-Nucleic acid), a key genetic material that stores crucial genetic information in cells. It is shown in Fig. 1.1.1. Nucleotides (deoxyribonucleotides) are the structural units of the DNA. Each nucleotide is composed of a pentose sugar (20-deoxy-D-ribose); one of the four nitrogenous bases—adenine (A), thymine (T), guanine (G), or cytosine (C); and a phosphate. Every base is attached to an atom of the sugar, and this forms a nucleoside, whereas nucleoside plus phosphate makes a nucleotide [14].

1.2 Central Dogma of Molecular Biology

The central dogma is the process of transforming genetic information from DNA to RNA to synthesize proteins. This process consists of DNA replication, coding for the RNA (mRNA- messenger RNA) through the transcription process, and then RNA codes final product proteins by translation as displayed in Fig. 1.2.1. Transcription is the process of passing the information from one strand of the DNA to RNA through complementary base pairing between DNA and the transcribed RNA. That is, an A in the DNA is transcribed to a U in the RNA, T to A, G to C.

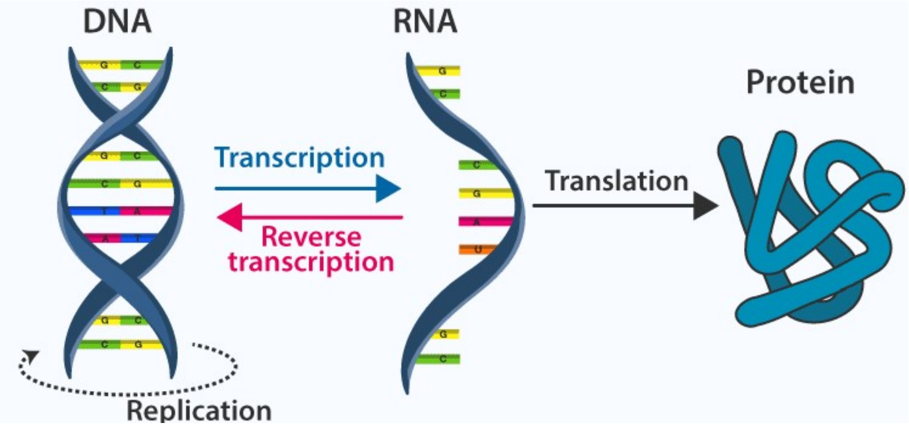

Fig. 1.2.1: Central dogma of molecular biology [5].

Proteins are chains of amino acids. There are 20 different standard amino acids that utilized in production of proteins. The translation process translates information from the language of nucleotides to that of amino acids. Then, parts of the mRNA are exported to the cytoplasm outside the nucleus, converted into protein. Transcribing genes to mRNAs, then converted into proteins, is called gene expression. And the abundance of a gene's mRNA molecules is typically called the expression value (level) of that gene or the gene expression [14].

1.3 Gene Expression

The fundamental physical and functional unit of heredity is a gene. The genes consist of DNA. Some genes serve as instructions for creating protein-called molecules. However, many genes do not code for proteins. Every person has two copies of each gene, one from each parent. Most genes are the same in all humans, but a small number of genes (less than 1 percent of the total) vary slightly between individuals [25]. Cell functions are determined by proteins. Thus, the thousands of genes expressed in a specific cell decide what the cell will do. Gene expression is the mechanism by which a cell reads the genetic code written in DNA to generate the molecule it requires. The cell interprets the genetic code to do this, and it adds one of the 20 different amino acids that are the basic units required to create proteins for each group of three letters [6].

The ability to monitor gene expression enables cells to have a functional protein if their normal functioning or survival needs it and this monitoring involves larger number of regulatory proteins. The regulation of gene expression conserves energy and space. It would take a large amount of energy for an organism to express every gene at all times, and so turning on the genes only when they are needed to save energy, and more importantly the combination of different genes that exhibit different functions. Each cell type in the body has a different set of active genes despite almost all body cells containing the same DNA. These different gene expression patterns cause your various cell types to have different sets of proteins, making each cell type uniquely specialized in doing its job.

1.4 RNA Sequencing

Next-generation sequencing (NGS) technologies are rapidly emerging and capable of processing large quantities of data at a lower cost and faster pace. When contrasted to their first-generation ancestor, the Sanger system, these innovations radically change researchers' sequencing capabilities [19]. And there are different types of sequencing methods as depicted such as DNA sequencing, RNA sequencing, and Methylation sequencing.

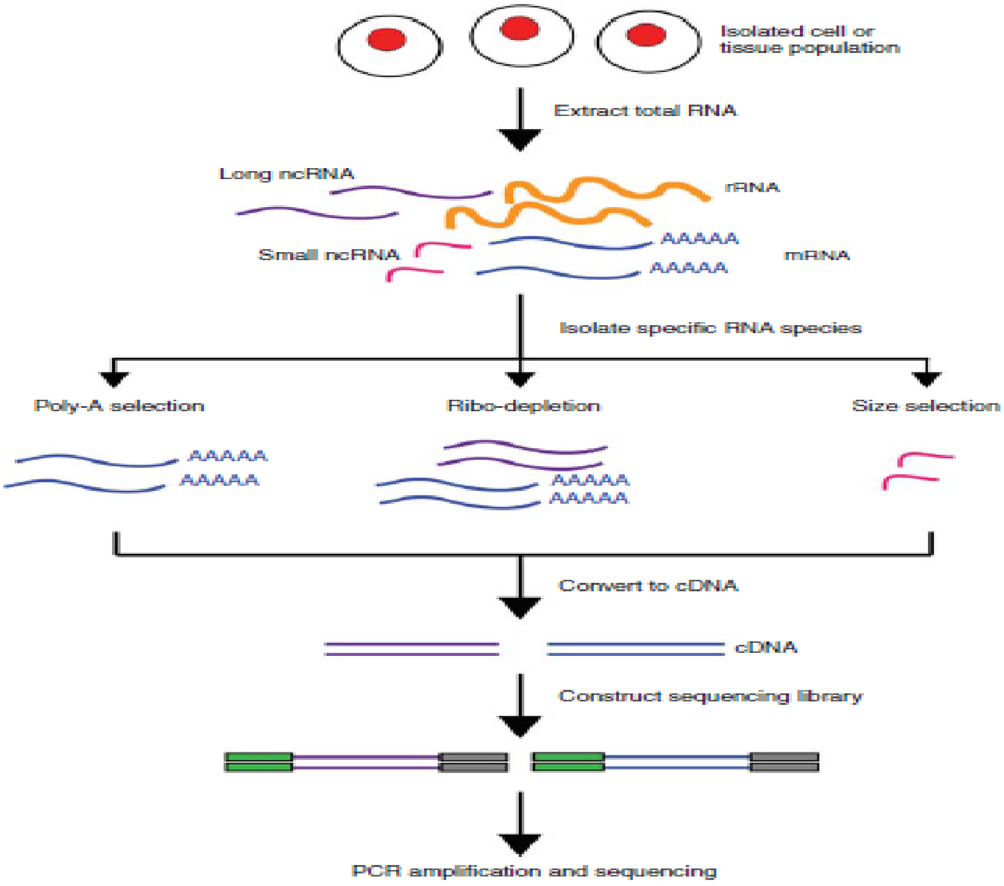

Fig. 1.4.1: Overview of RNA-seq workflow [15].

RNA-seq is a high-throughput transcriptome profiling technology alternative to the traditional RNA/cDNA cloning and sequencing strategies, and the brief overview of RNA-seq is shown in Fig.1.4.1. RNA-Seq transcripts are reverse-transcribed into cDNA and ligated to each end of the cDNA with adapters. Sequencing may be performed unidirectionally (single-end sequencing) or bidirectionally (paired-end sequencing). The findings can be matched to a reference genome database or assembled to generate de-novo transcripts, resulting in a genome-wide expression profile [13]. As a result, RNA-Seq offers many advantages over microarray technology. Microarray technology depends on already known genes, whereas RNA-Seq is not reliant on established genome data. It can screen new transcripts and examine transcript structure, including single base-pair resolution and exonic boundaries are extremely useful when examining SNPs (Single-nucleotide polymorphism), making it useful for genotyping and linkage analysis [13].

Through RNA-seq, many interesting biological experiments or discoveries are possible. Gene expression profiling between samples is one of the many scientific questions that RNA-seq can help with. The study of diseases linked to alternative splicing events (differential inclusion/exclusion of exons in the processed RNA product after splicing a precursor RNA segment).

1.5 Single-cell RNA sequencing (scRNA-seq)

Rapid advances in NGS technologies have provided many useful insights into complex biological processes, ranging from cancer genomics to diverse microbial species, in recent years. NGS-based genomics, transcriptomics, and epigenomics technologies are increasingly focusing on the characterization of single cells. Single cell sequencing analyses the sequence information from individual cells, allowing for a more thorough study of cellular variations and a deeper understanding of a cell's role in its micro-environment. Examining cells at the single-cell level offers provides opportunities to dissect the interplay between intrinsic cellular processes and extrinsic factors like the local environment or neighboring cells in cell fate determination [8]. Clinicians are also interested in single-cell studies because they can help them understand how an outlier cell can affect the outcome of an infection, drug or antibiotic resistance, and cancer relapse.

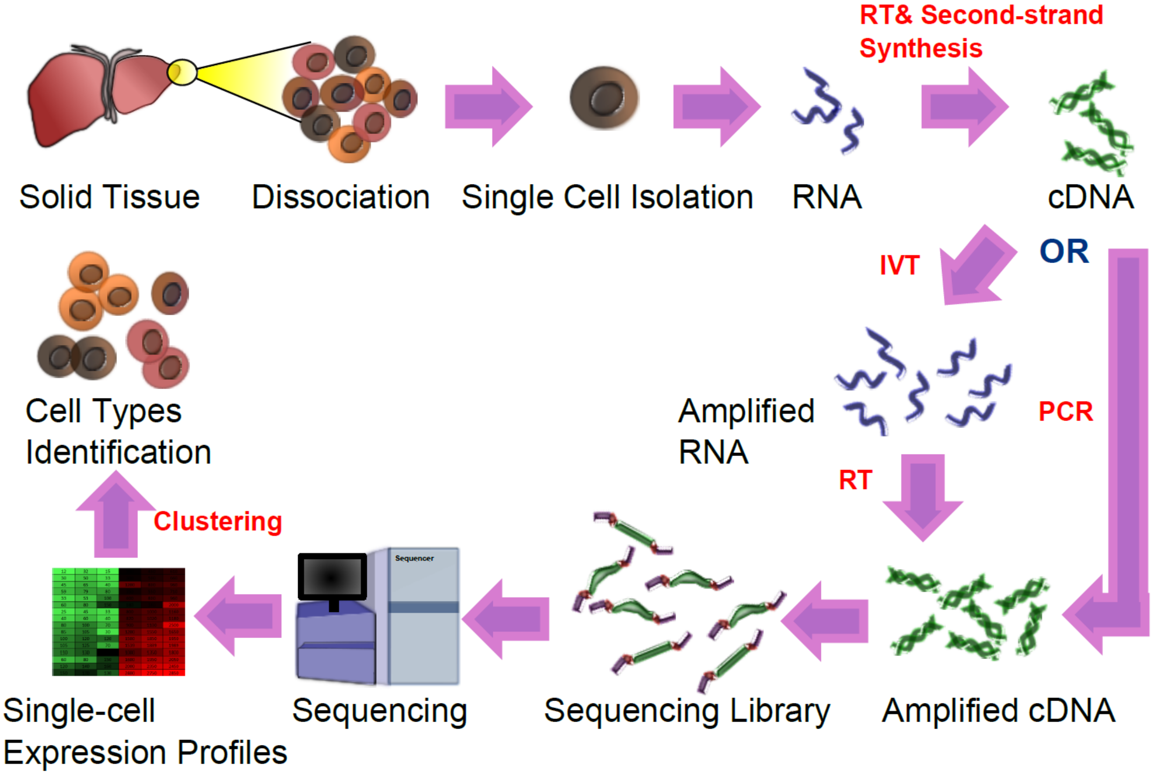

Fig. 1.5.1: Overview of scRNA-seq workflow [11].

Unknown organisms or regulatory processes of biotechnological or medical significance may also be discovered using scRNA-seq, and the overview of scRNA-seq library preparation is given in Fig.1.5.1. Global studies of single cells have been made possible by a substantial increase in the sensitivity of scientific instruments and the automation of all steps from sample preparation to data analysis. We can rapidly sequence the genomes of many single cells in parallel using next-generation sequencing techniques, or we can profile expressed proteins using fluorescence and mass cytometry. A variety of probe-dependent methods, such as reporter fusions to fluorescent proteins, fluorescence in-situ hybridization (FISH), quantitative real-time Polymerase chain reaction (RT-PCR), and microarrays, have pioneered mRNA profiling of single cells. Some of these methods can record expression changes of multiple genes in parallel. Trajectory inference, composition analysis, meta-stable states, cluster analysis, cluster annotation, gene expression dynamics, differential expression analysis, gene set analysis, and gene regulatory networks are just some of the applications that scRNA-seq can be used for [16].

1.6 Machine Learning

One way to explain the data and make predictions is to construct mathematical models. Another alternative is to use the data to build a prediction machine. This approach is known as machine learning, and it is a hot topic in the field of intelligent data processing. In bioinformatics, machine learning is commonly used [14]. In bioinformatics and genomics, for example, identifying genes and other functional elements on the genome is a challenging topic. The use of microarray data or proteomics expression data to diagnose cancers is a typical example. The gene expressions retrieved by microarrays form a vector for each patient. They can be thought of as the original features used to categorize the samples. A smaller number of genes may be selected to classify a particular type of cancer with normal cells or to classify subtypes of cancer. Traditionally, machine learning methods have been divided into three broad categories: supervised learning, in which the model is fed samples and predicted outcomes, unsupervised learning, in which the model is given unlabeled input samples, and reinforcement learning, in which the dynamic model where the agent takes the actions and it receives response from the environment it is in [14].

1.6.1 Unsupervised Learning

Unsupervised learning is when a machine learning model is given unlabeled data and asked to create some relationships between the data using various features. On the other hand, supervised learning teaches the model with current samples and their corresponding labels before predicting new samples. Cluster analysis is a standard unsupervised learning approach that groups data with many similarities into one or more groups.

1.6.2 Cluster Analysis

Cluster analysis is the job of grouping a collection of samples such that samples in one group (called a cluster) are more similar (in some idea) to those in other groups (clusters). For example, a person is sorting mixed grains from a bag and has no idea what they are or what they are used for. He must sort them into groups based on similarities such as shape, size, color, and so on. Here, mixed grains from the bag are referred to as unsupervised data. A person who does not have any knowledge but has to understand the data is unsupervised learning. Similarities in the machine learning model could be distance measures such as Euclidean, Manhattan, and Minkowski.

Hard and soft clustering are the two forms of clustering. Hard clustering involves placing any sample in the data into a cluster or not, while soft clustering involves assigning the cluster based on the possibility or probability of that point being in that cluster. Since clustering is a subjective process, there are various algorithms to accomplish it. For defining similarity across data points, each method has its own set of rules. In fact, there are over a hundred different clustering algorithms. Few widely used clustering techniques are k-means, spectral, expectation-maximization, hierarchical clustering. Cluster analysis has one more critical function. This is how the clustering efficiency is measured. The clusters' compactness and the distance between the clusters are two metrics to consider when evaluating clustering efficiency. Evaluating methods include the Silhoutte index [20], Calinski-Harabasz [3], and Davies Bouldin [9].

1.6.3 *k*-Means Algorithm

k-means is iterative clustering algorithm groups the data into n separate groups by minimizing the phenomenon within-cluster dispersion. The number of clusters/groups to be formed from the data needs to be specified as an input to the algorithm. The k-means algorithm is one of the widely used clustering algorithms since its inception.

$$SSE = \sum_{i=1}^{k} \min_{\mu \in C} (|\mathbf{x}_i - \mathbf{x}_j|)^2 \tag{1}$$

There are three major steps in k-means algorithm. The initial centroids are chosen in the first step. The most straightforward method is to choose k samples from the dataset X. k-means consists of looping between the two other steps after initialization. The first step is to allocate each sample to the centroid that is closest to it. The second step involves taking the mean value of all of the samples allocated to each previous centroid and creating new centroids. The algorithm computes the difference between the old and new centroids, then repeats the last two steps until the value is less than a threshold. In other words, it keeps repeating until the centroids do not move much (convergence). The points in the data choose such centroids where the compactness of the cluster is high or minimum sum of squared error (SSE) as given in (1) where n is number of samples in the data, C is cluster, μ is the mean of the samples, and x is corresponding sample. This algorithm always converges to local minimum, but not to the global minimum. There are many variants of k-means algorithms are present like minibatch k-means, and fuzzy k-means.

1.6.4 Dimensionality Reduction

The majority of real-life data, as well as machine learning data, is multidimensional. Furthermore, majority of the high-dimensional data is complex and sparse. Most importantly, understanding the data in such dimensions is difficult, and visualization is not possible. Data visualization is essential in a most of the machine learning tasks, and data reduction from higher dimensions to lower dimensions is needed. Dimensionality reduction is the process of transforming data from a high-dimensional space to a low-dimensional space while retaining some of the original data's meaningful properties, preferably close to its intrinsic dimension. Working in high-dimensional spaces may be inconvenient for various reasons: raw data is often sparse as a result of the curse of dimensionality, and data analysis is typically computationally intractable. Dimensionality reduction is divided into categories such as linear and non-linear. Few of the mainly used techniques are listed below.

1. Principal Component Analysis (PCA): PCA is a popular linear technique for

dimensionality reduction. Given a set of data with n dimensions, PCA aims to find a linear subspace of dimension d lower than n such that the data points lie mainly on this linear subspace. Such a reduced subspace attempts to maintain variability of the data [12].

2. t-distributed Stochastic Neighbor Embedding (t-SNE): t-SNE is non-linear dimensionality reduction technique used for visualizing high-dimensional datasets and gives decent visualizations in lower spaces. t-SNE is not used for cluster analysis or outlier detection since it does not preserve the data's distances or densities. It is extensively applied in image processing, Natural Language Processing, genomic data and speech processing [23].

3. Autoencoders, a deep learning technique can also be used for dimensionality reduction with an inverse function from the coding to the original representation [24].

4. Uniform Manifold Approximation and Projection (UMAP): a strategy for reducing dimensionality that is non-linear. It resembles t-SNE in appearance, but it assumes that the data is uniformly distributed on a locally connected Riemannian manifold and that the Riemannian metric is locally constant or roughly locally constant [18].

There are a few other non-linear dimensionality reduction techniques referred to as manifolding techniques and these techniques produce a compact low dimensional embeddings. Given, $X = x_1, x_2, ..., x_n \in \mathbb{R}^D$ and we want to reduce the data from higher dimensions to lower dimensions. The data lies in d dimensions embedded into $\mathbb{R}^D$, where $d < D$. The aim is to learn a manifold from a set of points. Oft-used and straightforward examples in the manifold learning literature are the S-shape, Swiss roll, and circular shape two-dimensional manifold embedded in $\mathbb{R}^3$. Figure 1.6.1 shows the S-shape and a learned two-dimensional embedding of shapes found using manifolding techniques[4]. These methods begin by constructing a sparse graph in which the nodes represent input patterns, and the edges represent neighborhood relations.

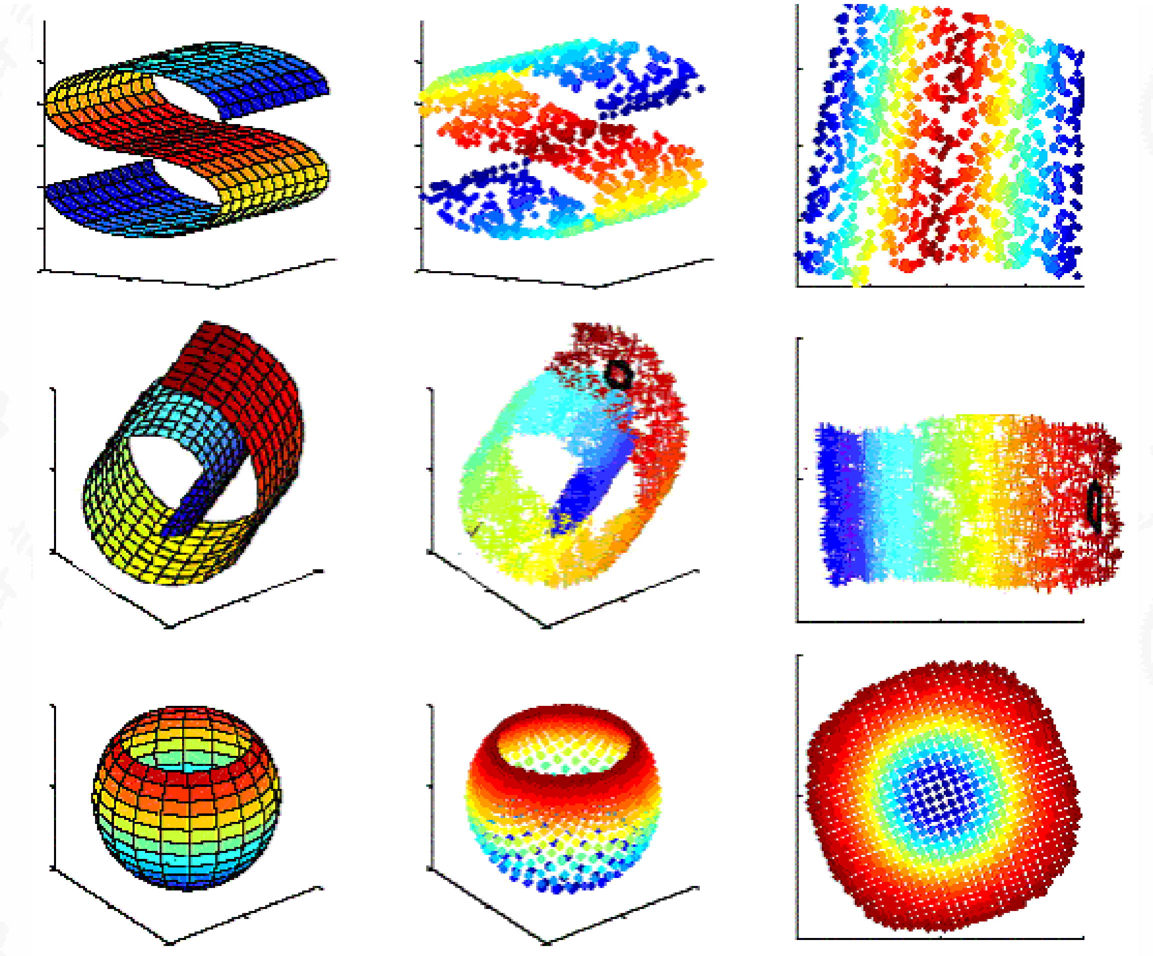

Fig. 1.6.1: Dimensionality reduction from three dimensions to two dimensions using manifold techniques [17].

Suppose that the resulting graph is connected and can be viewed as an approximation of the submanifold sampled by the input patterns. From these graphs, one can then construct matrices whose spectral decomposition reveal the submanifold's low dimensional structure (and sometimes even the dimensionality itself). Also, these techniques preserve the relationships of the samples from high dimensions to lower dimensions through geodesic distances among the samples. The shortest path calculates the geodesic distance to the nearest neighbor from a given sample[22].

1. Isomap [1] works by computing the low-dimensional representation of a high-dimensional data set that preserves the pairwise distances between input patterns measured along the geodesic submanifold from which they were sampled.

2. Locally Linear Embedding (LLE): LLE [21] is focused on finding the lowest-dimensional representation of high-dimensional data set that preserves the local linear structure of neighbouring input patterns the most accurately. The outputs of the algorithm are extracted from the smallest eigenvectors of a sparse matrix, rather than the largest eigenvectors of a (dense) Gram matrix, which distinguishes it from Isomap and maximum variance unfolding.

3. Laplacian Eigenmaps: The structure of this algorithm is close to that of LLE. Laplacian eigenmaps map nearby input patterns to nearby outputs by computing the low-dimensional representation of a high-dimensional data set that most faithfully preserves proximity relations [2].

4. Multi-dimensional Scaling (MDS): MDS [7] is another traditional method for mapping high-dimensional data onto a lower-dimensional space when attempting to preserve pairwise distances among data points. That is, MDS uses knowledge about the distances between the patterns to solve the problem of constructing a configuration of points in the Euclidean space.

Dimensionality reduction can be used for data visualization, noise reduction, cluster analysis, or intermediate steps to assist other steps in data analysis.

1.7 Motivation

scRNA-seq data occurs in high dimensions and it is usually sparse. Barcodes with a low count depth, a few detectable genes, and a high fraction of mitochondrial counts indicate that cytoplasmic mRNA has leaked out via a broken membrane, leaving only mitochondrial mRNA to be conserved. Cells with unusually high counts and many detectable genes, on the other hand, could be doublets. In total, the data has a significant amount of unwanted information such as low-quality cells, ERCC spike-ins, and mitochondrial genes that do not contribute to the downstream analysis. As a result, pre-processing the data is very important. Cells, on the other hand, use morphological properties, position, and form to describe themselves. None of these properties carry enough information to result, in every case, in a accurate cell type identification. Identifying relevant components, such as target cells, is a critical step in characterizing diseases, leading to enhanced diagnosis, treatment, and prognosis. In particular, analyzing the variety and evolution of single cancer cells can help in early cancer detection and, eventually, select the best cancer treatment plan. As a result, the problem of recognizing cell types, cell groups, or cell sub-populations in single-cell data becomes very interesting.

1.8 Problem Statement

Given the complex, high-dimensional scRNA-seq data, the problem is made more complicated because some of the data is heterogeneous. Unsupervised learning is the method of choice when researchers have no prior knowledge of the results. Given single-cell RNA data, the aim is to identify potential marker genes, create a pipeline that analyzes single-cell data, and cell types by grouping the genes into different clusters using machine learning techniques such as dimensionality reduction, clustering, and statistical ranking.

1.9 Proposed Method

This thesis proposes a new pipeline for identifying cell types in single-cell data that uses a hybrid model of non-linear dimensionality reduction techniques, linear combination methods for visualization, and clustering using k-means. Higher-dimensional data is reduced to a smaller number of dimensions. The goal is to identify cell types from single-cell RNA-Seq data, which could be used to understand diseases and help identify drugs and study types of treatments needed for a specific disease.

Before applying machine learning techniques, the primary step is to format the data for downstream analysis. The pre-processing step includes filtering out low-quality cells and removing unwanted genes. Normalization is used to transform the data without altering its context since the data is genomic and presented at various expression levels. As part of the feature selection process, highly variable genes are extracted, and then those genes are passed to further steps. For dimensionality reduction, combining non-linear and linear approaches results in better lower embeddings and visualizations that can be clustered efficiently. Clusters are evaluated using validity indices, obtaining the highest possible score. Once the clusters are formed, the top marker genes from each cluster are extracted, and those marker genes are used to search biological databases for cell types. The cell type with the largest number of overlapping genes is selected and annotated to the cluster that corresponds to it.

1.9.1 Contributions

This thesis introduces a new pipeline that can be applied to any scRNA-seq data to discover cell types, identify marker genes and provide enhanced visualization. This pipeline employs machine learning algorithms such as MLLE and ICA for dimensionality reduction (to escape from the curse of dimensionality) and visualization, respectively. The data is clustered, top marker genes are extracted using Wilcoxon method, a gene ranking technique, and cell types are discovered; these results can be used to study diseases. The contributions of this thesis can be summarized as follows:

- Proposed a new validated pipeline to identify cell types in single-cell RNA-seq data.
- Proposed a new combination and a hybrid model for dimensionality reduction using linear and non-linear methods, which is powerful in exploring the data in lower dimensions.
- Proposed a mechanism to choose the optimized combination of the number of nearest neighbors and the number of clusters k for clustering and dimensionality reduction using validity indices.
- Developed an open-source software tool for identifying cell types in single-cell data, available at GitHub project.

References

[1] Mukund Balasubramanian et al. “The isomap algorithm and topological stability”. In: *Science* 295.5552 (2002), pp. 7–7.

[2] Mikhail Belkin and Partha Niyogi. “Laplacian eigenmaps for dimensionality reduction and data representation”. In: *Neural computation* 15.6 (2003), pp. 1373–1396.

[3] Tadeusz Caliński and Jerzy Harabasz. “A dendrite method for cluster analysis”. In: *Communications in Statistics-theory and Methods* 3.1 (1974), pp. 1–27.

[4] Lawrence Cayton. “Algorithms for manifold learning”. In: *Univ. of California at San Diego Tech. Rep* 12.1-17 (2005), p. 1.

[5] *Central Dogma - Steps Involved in Central Dogma.* en-US. URL: `https://byjus.com/biology/central-dogma-inheritance-mechanism/` (visited on 04/02/2021).

[6] Supratim Choudhuri. *Bioinformatics for beginners: genes, genomes, molecular evolution, databases and analytical tools.* Elsevier, 2014.

[7] Michael AA Cox and Trevor F Cox. “Multidimensional scaling”. In: *Handbook of data visualization.* Springer, 2008, pp. 315–347.

[8] Saiteja Danda, Akram Vasighizaker, and Luis Rueda. “Unsupervised Identification of SARS-CoV-2 Target Cell Groups via Nonlinear Dimensionality Reduction on Single-cell RNA-Seq Data”. In: *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM).* IEEE. 2020, pp. 2737–2744.

[9] David L Davies and Donald W Bouldin. “A cluster separation measure”. In: *IEEE transactions on pattern analysis and machine intelligence* 2 (1979), pp. 224–227.

[10] *DNA structure. Illustration of helix, cell - 56366025.* en. URL: `https://www.dreamstime.com/stock-illustration-dna-structure-nucleotide-phosphate-sugar-bases-image56366025` (visited on 04/01/2021).

[11] *English: Single cell RNA sequencing workflow.* Feb. 2014. URL: `https://commons.wikimedia.org/wiki/File:Single_cell_RNA-Seq_workflow.pdf` (visited on 04/02/2021).

[12] Ali Ghodsi. "Dimensionality reduction a short tutorial". In: *Department of Statistics and Actuarial Science, Univ. of Waterloo, Ontario, Canada* 37.38 (2006), p. 2006.

[13] "Chapter 18 - Genomic and Proteomic Mechanisms and Models in Toxicity and Safety Evaluation of Nutraceuticals". In: *Nutraceuticals.* Ed. by Ramesh C. Gupta. Academic Press, 2016.

[14] Rui Jiang, Xuegong Zhang, and Michael Q Zhang. *Basics of bioinformatics: Lecture notes of the graduate summer school on bioinformatics of China.* Springer Science & Business Media, 2013.

[15] Kimberly R Kukurba and Stephen B Montgomery. "RNA sequencing and analysis". In: *Cold Spring Harbor Protocols* 2015.11 (2015), pdb–top084970.

[16] Malte D Luecken and Fabian J Theis. "Current best practices in single-cell RNA-seq analysis: a tutorial". In: *Molecular systems biology* 15.6 (2019), e8746.

[17] *Manifolding Techniques.* en-US. URL: `http://1.bp.blogspot.com/-OY4AbFjqAbY/VSQZTLHSlKI/AAAAAAAAApQ/E2RaiIpCwQQ/` (visited on 04/03/2021).

[18] Leland McInnes, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction". In: *arXiv preprint arXiv:1802.03426* (2018).

[19] "Next Generation Sequencing". In: *Pathobiology of Human Disease.* Ed. by Linda M. McManus and Richard N. Mitchell. Academic Press, 2014.

[20] Peter J Rousseeuw. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". In: *Journal of computational and applied mathematics* 20 (1987), pp. 53–65.

[21] Sam T Roweis and Lawrence K Saul. "Nonlinear dimensionality reduction by locally linear embedding". In: *science* 290.5500 (2000), pp. 2323–2326.

[22] Lawrence K Saul et al. "Spectral methods for dimensionality reduction." In: *Semi-supervised learning* 3 (2006).

[23] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11 (2008).

[24] Wei Wang et al. "Generalized autoencoder: A neural network framework for dimensionality reduction". In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2014, pp. 490–497.

[25] *What is a gene?: MedlinePlus Genetics*. en. URL: `https://medlineplus.gov/genetics/understanding/basics/gene/` (visited on 04/04/2021).

CHAPTER 2

Discovering Cell Types Using Manifold Learning and Enhanced Visualization of Single-cell RNA-Seq Data

2.1 Introduction

Single-cell sequencing is an emerging technology used to capture cell information at a single-nucleotide resolution and by which individual cells can be analyzed separately [13]. As of now, single-cell RNA-seq (scRNA-seq) datasets have been generated for different purposes [15]. However, these high-dimensional and sparse data lead to some analytical challenges. While many computational methods have been successfully proposed for analyzing scRNA-seq data, there are still some open problems in this research area. One of the main challenges is sparsity of data and the curse of dimensionality presented in scRNA-seq data. Also, performing well-defined pre-processing steps leads to enhance the quality of data and new biological insights. Analyzing scRNA-seq data can be divided into two main categories: at the cell level and gene level. Finding cell sub-types or highly differentially expressed tissue-specific gene set is one of the common challenges at the cell level [27]. Arranging cells into clusters to find the data's heterogeneity is arguably the most significant step of any scRNA-seq data downstream analysis. This step could be used to distinguish tissue-specific subtypes based on identified gene sets. Indeed, cell clustering aims to identify cell types based on the patterns embedded in gene expression without prior knowledge at the cell level. Since the number of genes that are profiled in scRNA-seq data is typically large, cells tend to be located close to each other via non-metric distances, but rather complex relationships in high-dimensional spaces [20]. Therefore, traditional dimensionality reduction and clustering algorithms are unsuitable for these scenarios, and hence, they cannot efficiently separate individual cell types. Several algorithms have been proposed to lower the dimension of the data and cluster cells from scRNA-seq profiles to alleviate the problem of curse of dimensionality.

Dimensionality reduction techniques have been widely used in several studies of large-scale scRNA-seq data processing [7]. Most of the previous studies use principal component analysis (PCA). However, one of the main drawbacks of PCA is that it cannot deal with sparse matrices and non-metric relationships among high-dimensional data points. Also, there was no advantage in keeping the clustering performance after the changes in the data in lower dimensions [9]. Other works have also employed PCA as a pre-processing step to remove cell outliers for performing dimensionality reduction and visualization. Other methods proposed nonlinear dimensionality reduction methods such as t- distributed Stochastic Neighborhood Embedding (t-SNE) [33] and UMAP [23]. However, UMAP and t-SNE is not useful for high-dimensional cytometry. Moreover, several studies have used unsupervised clustering models to identify rare novel cell types. For instance, the hierarchical clustering algorithm divides large clusters into smaller ones or merge each data points into larger clusters progressively. This algorithm has been employed to analyze scRNA-seq data by BackSPIN [40] and pcaReduce [39], through dimension reduction after each division or combination in an iterative manner. k-means, which is one of the most common clustering algorithms has been employed in the Monocle, specifically for analyzing scRNA-seq data [25]. Also, the authors of [36] used the Louvain algorithm, which is based on community detection techniques to analyze complex networks [14].

However, to achieve acceptable clustering performance on scRNA-seq data, other comprehensive studies indicated that hybrid models, designed as a combination of clustering and dimensionality reduction techniques, tend to improve the clustering results [9]. They learned 20 different models using four dimensionality reduction methods, including PCA, non-negative matrix factorization (NMF), filter-based feature selection (FBFS), and Independent Component Analysis (ICA). They also used five clustering algorithms as k-means, density-based spatial clustering of applications with noise (DBSCAN), fuzzy c-means, Louvain, and hierarchical clustering. Their experiments highlighted the positive effect of hybrid models and showed that using feature-extraction methods could be a decent way to improve clustering performance. Their experimental results indicate that Louvain combined with ICA performed well in small feature spaces.

In this paper, we proposed a model to obtain efficient and meaningful clusters of cells from large-scale scRNA-seq data. We focus on the combination of unsupervised dimensionality reduction followed by conventional clustering. We discovered a hybrid model of non-linear dimensionality reduction technique (MLLE) and linear combination method (ICA) for visualization and compared it to PCA, t-SNE, Isomap, regular Locally Linear Embedding (LLE), and Laplacian eigenmaps. ICA is employed to enhance visualization and clustering of the data. Parameter tuning or choosing the best parameters for dimensionality reduction and clustering has been one of the main challenges in the field that is well addressed in our work. Experimental results on thirteen different benchmark scRNA-seq datasets show the power of modified LLE and ICA on clustering data and representation quality, providing very high accuracy and enhanced visualization. Confirmatory biological annotations were observed in the clusters using corresponding marker genes found by our method.

2.2 Materials and Methods

The block diagram of the proposed pipeline is depicted in Fig. 2.3.1. The scRNA-seq data is first pre-processed based on the number of cells and the number of genes obtained in the first step. Highly variable genes are extracted as part of the feature selection step after normalization and scaling of the filtered data. Linear regression is one of the most widely-used methods to correct potential technical batch effect present in the data based on the total counts per cell and mitochondrial percentage as discussed in [36] [22]. The data obtained at this point is then processed to reduce the dimensions of the feature space into two or three dimensions; afterward, k-means clustering is applied. In addition, we performed ICA on the lower-dimensional data followed by k-means clustering to achieve meaningful clusters and enhanced visualization.

2.3 Materials and Methods

The block diagram of the proposed pipeline is depicted in Fig. 2.3.1. The scRNA-seq data is first pre-processed based on the number of cells and the number of genes obtained in the first step. Highly variable genes are extracted as part of the feature selection step after normalization and scaling of the filtered data. Linear regression is one of the most widely-used methods to correct potential technical batch effect present in the data based on the total counts per cell and mitochondrial percentage as discussed in [36] [22]. The data obtained at this point is then processed to reduce the dimensions of the feature space into two or three dimensions; afterward, k-means clustering is applied. In addition, we performed ICA on the lower-dimensional data followed by k-means clustering to achieve meaningful clusters and enhanced visualization.

2.3.1 Datasets

To evaluate the performance of the proposed method, a total of thirteen benchmark scRNA-seq datasets were used, which include single-cell gene expression profiles. The details of all datasets used in this work are given in Table 2.3.1. They vary across size, tissue (pancreas, lung, peripheral blood), sequencing protocol (three different protocols), and species (Human and Mouse). Datasets Xin[38], H1299_scRNAseq[37], and Calu3_scRNAseq[37] datasets are unlabeled and do not have any background knowledge of the data. In this case, we analyzed the data and provided useful information

Table 2.3.1: Datasets used in this work.

Dataset	No. of cells	No. of genes	Accession number	Description	Sequencing technology
Baron_human1	16,381	1,937	GSE84133	Human pancreas	llumina HiSeq 2500 (inDrop)
Baron_human2	16,381	1,724	GSE84133	Human pancreas	llumina HiSeq 2500(inDrop)
Baron_human3	16,381	3,605	GSE84133	Human pancreas	llumina HiSeq 2500(inDrop)
Baron_human4	16,381	1,303	GSE84133	Human pancreas	llumina HiSeq 2500(inDrop)
Baron_mouse1	14,878	822	GSE84133	Mouse pancreas	llumina HiSeq 2500(inDrop)
Baron_mouse2	14,878	1,064	GSE84133	Mouse pancreas	llumina HiSeq 2500(inDrop)
Muraro	17,140	3,071	GSE85241	Human Pancreas	Illumina NextSeq 500 (CEL-Seq2)
Segerstolpe	26,271	7,028	E_MTAB_5061	Human Pancreas	Smart-Seq2
Xin	39,851	1,601	GSE81608	Human Pancreas	Illumina HiSeq 2500(SMARTer)
Wang	19,950	635	GSE83139	Human Pancreas	Illumina HiSeq 2000(SMARTer)
H1299_scRNAseq	48,890	27,072	GSE148729	Human lung (SARS-CoV-2)	Illumina NextSeq 500
Calu3_scRNAseq	24,754	27,072	GSE148729	Human lung (SARS-CoV-2)	Illumina NextSeq 500
PBMC	32,738	2,700	10X Genomics (pbmc3k)	3k PBMCs from a Healthy Donor	Cell Ranger

about the unknown data. On the other hand, pancreas datasets including Baron [1] , Muraro [24], Segerstolpe [29], Xin [38], and Wang[35]. Moreover, peripheral blood dataset, 3k PBMC from a healthy donor, were downloaded from the 10XGenomics portal[10]. H1299_scRNAseq and Calu3_scRNAseq datasets (GSE148729) were extracted from NCBI's Gene Expression Omnibus [32].

2.3.2 Data Pre-processing and Quality Control

A common practice for generating RNA-seq raw data is to use next-generation sequencing technologies to create read count matrices. The read count data matrix contains gene names and their expression levels across individual cells. Before analyzing scRNA-seq data, one needs to ensure that gene expressions and cells are of standard quality. We follow a typical scRNA-seq analysis workflow including quality control, as described in [22][18]. Based on the expression levels, we filtered out weakly expressed genes and low-quality cells in which fewer reads are mapped, as shown in Fig. 2.3.1, the first step of pre-processing. Low-quality cells that are dyed, degraded, or damaged during sequencing are represented by a low number of expressed genes. Genes expressed in less than three cells and cells with less than 200 expressed genes are removed. This step is performed to remove low quality cells and poorly expressed

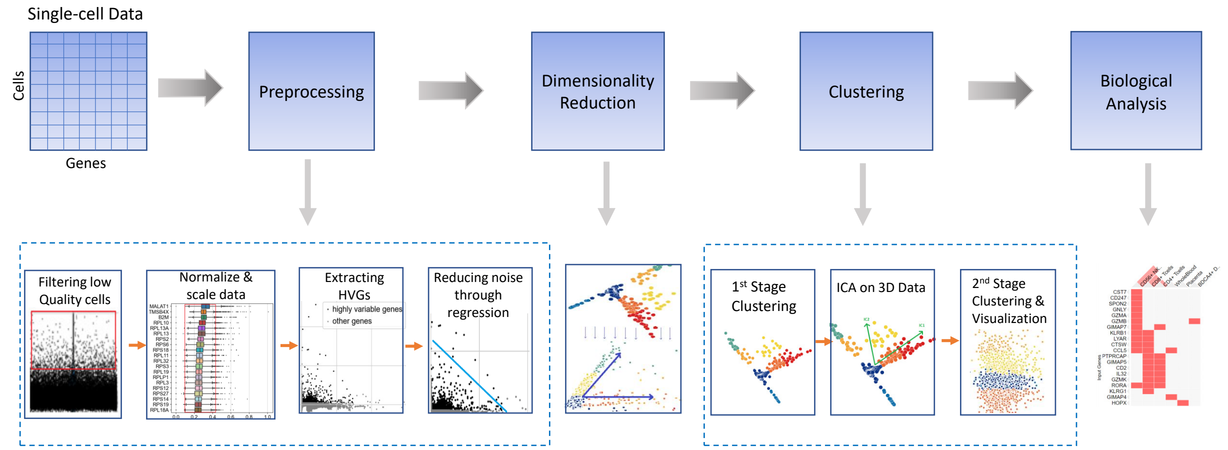

Fig. 2.3.1: Block diagram of the proposed approach for discovering cell types in scRNA-seq data.

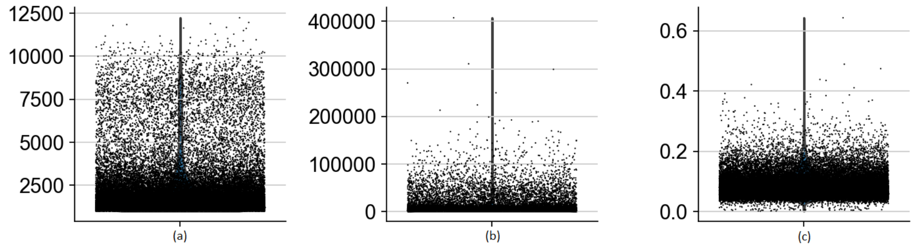

Fig. 2.3.2: Investigating the distribution of the data to filtered out weakly expressed genes and low-quality cells from dataset; (a) number of expressed genes, (b) total counts per cell, and (c) the percentage of mitochondrial genes for H1299_scRNAseq.

genes.

We also investigated the distribution of the data (Fig.2.3.2) as a data-specific quality-control step and filtered out low-quality cells and genes. Also, we remove a percentage of mitochondrial genes that do not contribute significant information to the downstream analysis [19], [18].

Since the scRNA-seq data expressed at different levels, normalization is a must. Normalization is the method of translating numeric columns' values in a dataset to a standard scale without distorting the ranges of values. Visualization of top genes in the dataset are shown in Figures 2.3.3 and 2.3.4 before and after normalization, respectively. We normalize the data using the Counts Per Million (CPM) normalization

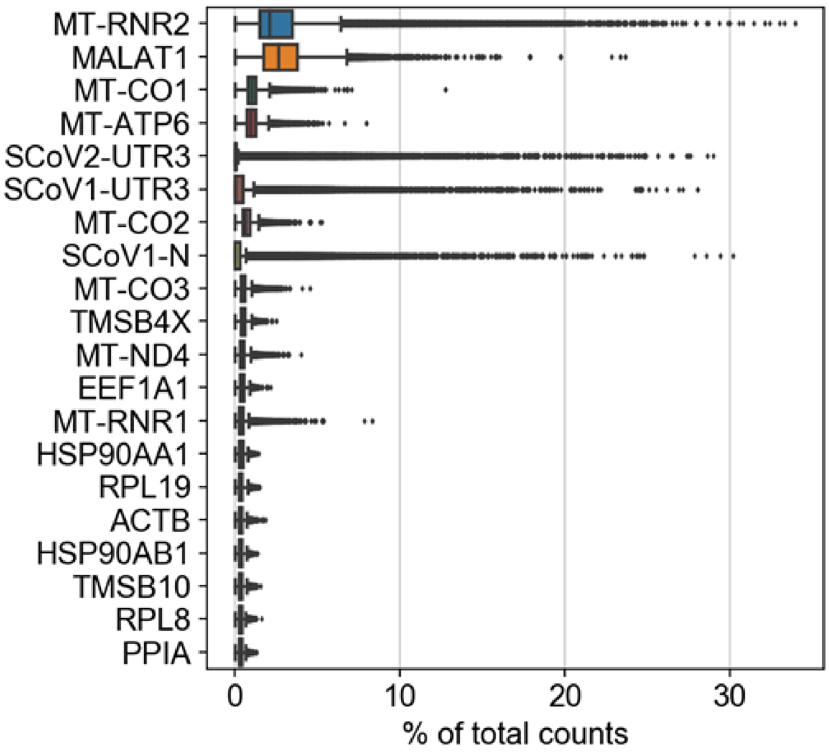

Fig. 2.3.3: Top 20 highly-variable genes before normalization.

combined with logarithmic scaling on the data:

$$CPM = readsMappedToGene \times \frac{1}{totalReads} \times 10^6 \tag{1}$$

where $totalReads$ is the total number of mapped reads of a sample, and $readsMappedToGene$ is the number of reads mapped to a selected gene.

At this point, we extracted highly variable genes (HVGs) as a part of the feature selection step, aiming at minimizing the search space, and only these genes are examined in further evaluation. We then removed any random noise and held genes that highlight relevant biological information. HVGs are those genes that are expressed significantly more or less in some cells compared to other ones. This step in quality control makes sure that the differences occur because of biological differences and not technical noise. The simplest approach to compute such a variation is to quantify the variance of the expression values for each gene across all samples. A good trade-off

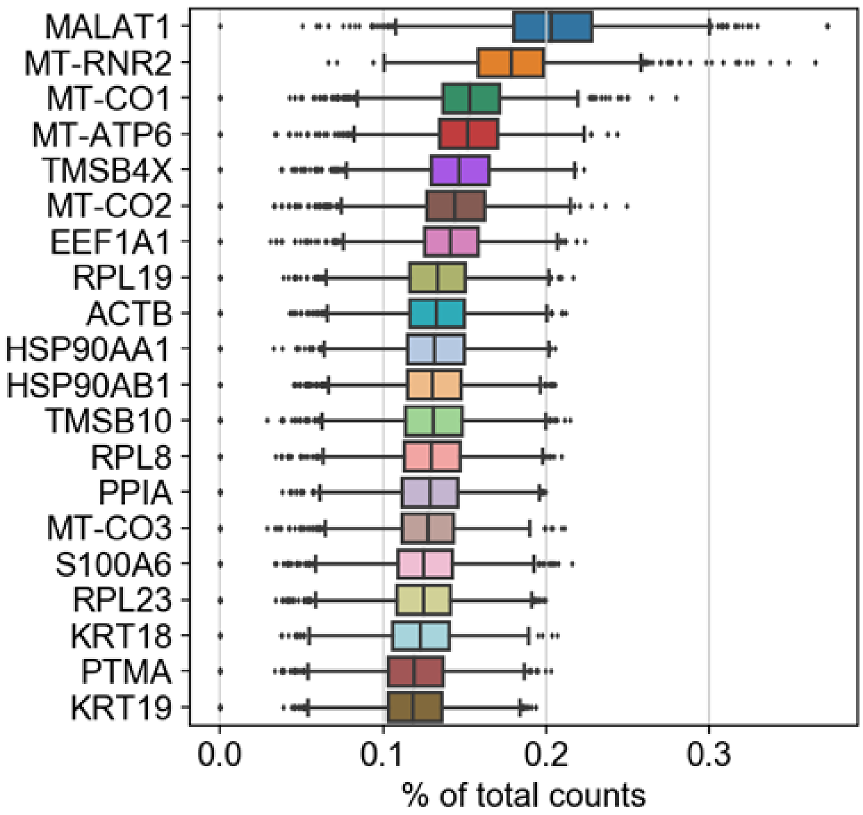

Fig. 2.3.4: Top 20 highly-variable genes after normalization.

between mean and variance would help select the subset of genes that keep useful biological knowledge, while removing noise. We use log-normalized data because we want to ensure having the same log-values in the clustering and dimensionality reduction follow a consistent analysis through all steps. There are several widely-used approaches to find the best threshold. The normalized dispersion is obtained by scaling the mean and standard deviation of the dispersion for genes falling into a given bin for the mean expression of genes (Fig. 2.3.7). This means that for each bin of mean expression, HVGs are selected. A Python package, Scanpy, is used to perform pre-processing and quality control steps.

2.3.3 Dimensionality Reduction

The majority of real-life data is multidimensional. Furthermore, the majority of the high-dimensional data is complex and sparse. Most importantly, understanding the

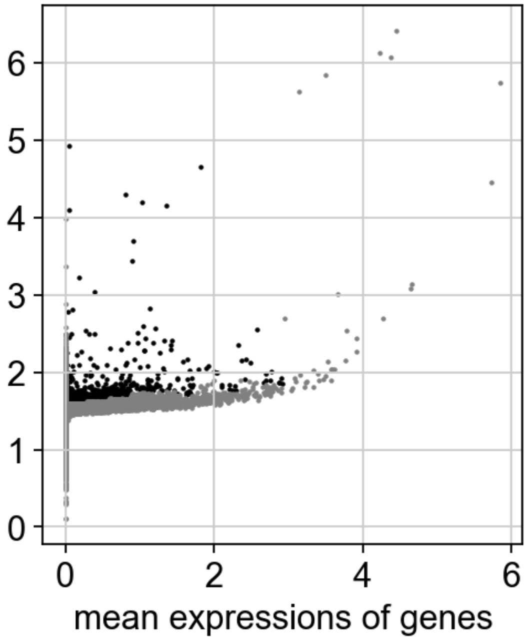

Fig. 2.3.5: Dispersion of genes before normalization.

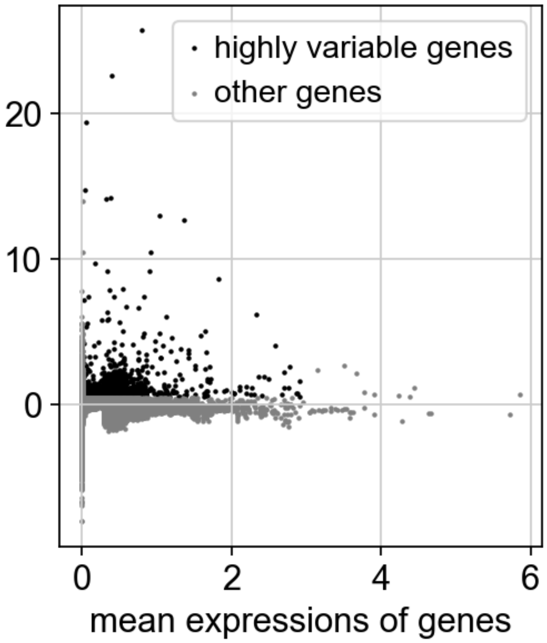

Fig. 2.3.6: Dispersion of genes after normalization.

Fig. 2.3.7: Comparison of dispersion of normalized and not normalized genes to extract highly variable genes.

data in such dimensions is tricky, and visualization is not possible. Dimensionality reduction is the process of transforming data from a high-dimensional space to a low-dimensional space while retaining some of the original data's meaningful properties, preferably close to its intrinsic dimension. Working in high-dimensional spaces may be inconvenient for various reasons: raw data is often sparse as a result of the curse of dimensionality, and data analysis is typically computationally intractable. On the other hand, high-dimensional gene expression data is complex and should be well-explored. Each gene is characterized as a data dimension in a single-cell expression profile in a single-cell expression profile. As such, dimensionality reduction is very productive in summarizing biological attributes in fewer dimensions. Dimensionality reduction is divided into linear and non-linear techniques.

2.3.3.1 Modified Locally Linear Embedding

MLLE is the enhanced version of LLE and hence the authors named it as Modified LLE. To understand the working of MLLE, we need to understand LLE. LLE tries to reveal the manifold's underlying structure based on simple geometric intuitions when used for dimensionality reduction. LLE preserves the data's locality in lower dimensions because it reconstructs each sample point from its neighbors. In the simplest formulation of LLE, one identifies nearest neighbors per data point, as measured by Euclidean distance[28]. One can choose number of neighbors based on some rules or using some metrics or some random number. Consider the sample points $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ in high dimensional space, where $\{\mathbf{x}_j, j \in N\}$ and $\mathbf{W} = \{w_{ij}\}$ is the weight matrix. A directed graph $\mathcal{G} = (\mathbf{X}, \mathbf{E}, \mathbf{W})$ is constructed considering the neighborhood relations of the sample points $\mathbf{X}$, in high dimensional space, and $\mathbf{E} = \{e_{ij}\}$ represents the edges of the graph. Later, weights are assigned to the edge of the graph. To compute the weights $\mathbf{W_{kn}}$, minimize the cost function with respect to two constraints: 1) each data points $\mathbf{x_i}$, is reconstructed only from its neighbors imposing $\mathbf{W_{kn}} = \mathbf{0}$ if $\mathbf{x_i}$ does not belong to that set, 2) sum of the weights matrix rows equal to one, that is $\mathbf{W_{kn}} = 1$. Optimal weights are calculated by solving (2) the constrained squared distances problem shown below [28].

$$\min \ \mathbf{x}_i - \sum_{k\in K_n} w_{kn}\mathbf{x}_k \quad \text{s.t.} \sum_{k\in K_n} w_{kn} = 1\,. \tag{2}$$

The computed weights are then allocated to each edge of the graph, with each data point viewed as a small linear patch of the sub-manifold.

Finally, each high-dimensional input sample $\mathbf{x_i}$ mapped to a low dimensional point set $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n\}$ representing the manifold's global internal coordinates. The reconstruction weights for each data point are calculated independently of the weights for other data points from its local neighborhood. The embedding coordinates are computed by an NXN eigen solver, a global operation that combines all data points in connected components of the graph identified by the weight matrix. While reconstructing the structure from the higher dimension to the lower dimension, some information could be lost. This lost information is noted as a reconstruction error and computed using (3).

$$\epsilon_r = \sum_{i=1}^{n} |\mathbf{y}_i - \sum_{k\in K_i} w_{ik}\mathbf{y}_k|^2 \tag{3}$$

The regularization problem is a well-known issue with LLE. The matrix representing each local neighborhood is rank-deficient when the number of neighbors exceeds the number of input dimensions. To deal with this, standard LLE uses an arbitrary regularisation parameter in relation to the weight matrix's local trace[**mlle**]. This problem manifests itself in embedding which distort the underlying geometry of the manifold. MLLE is one such technique, which overcomes this regularization problem using multiple weights in each neighborhood. MLLE modifies or adjusts the reconstruction weights [8] shown in (2) and this modifies the objective function (3).

$$\epsilon_r = \sum_{i=1}^{n} \sum_{l=1}^{s_i} |\mathbf{y}_i - \sum_{k\in K_i}^{w} {}_{ik}\mathbf{y}_k|^2 \tag{4}$$

where, s_i = smallest right singular vectors of $\mathbf{G}$.

MLLE aims to take advantage of the dense relations that exist in the embedding space. It is closely related to the other version of the LLE, that is Local Tangent Space Alignment (LTSA) [34].

2.3.3.2 Independent Component Analysis

ICA is an independent and linear dimensionality reduction method. By using simple statistical properties assumptions, ICA learns an efficient linear transformation of the data and attempts to find the underlying components and sources present in the data [16]. Unlike other approaches, the transformation's underlying vectors are presumed to be independent of one another. It employs a non-Gaussian data structure, which is crucial for retrieving the transformed underlying data components. Consider, $\mathbf{r}$ is a random vector whose elements are $\{\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_n\}$, and similarly, random vector $\mathbf{s}$ with its elements $\{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n\}$, and $\mathbf{A}$ is the matrix with elements a_{ij}. The ICA model is a generative model, and it explains how the observed data are generated (5) by mixing the components s_i. The independent components are latent variables, which means they are unknown. Also, the mixing matrix is assumed to be unknown.

$$\begin{aligned} \mathbf{r} &= \mathbf{As} \\ \mathbf{Y} &= \mathbf{AX} \end{aligned} \tag{5}$$

Rows of these vectors and the matrix are orthogonal to each other. As such, it leads to more informative components than PCA. ICA does not require knowing the system's output to break the data into some measurements. Hence it is referred to as blind source separation[17]. Here, a source means the original data, independent components. Blind means that it knows nothing but very little, if anything, on the mixing matrix and makes modest assumptions on the source data.

2.3.3.3 Other Dimensionality Reduction Methods

We used other dimensionality reduction techniques to compare our proposed method such as Standard LLE, Isomap, Laplacian eigenmap, PCA, and t-SNE. Isomap stands for isometric mapping. Isomap is a non-linear dimensionality reduction method based on the spectral theory that tries to preserve the lower dimension's geodesic distances. Isomap starts by creating a neighborhood network. After that, it uses graph distance to estimate the geodesic distance between all pairs of points. The eigenvalue decomposition of the geodesic distance matrix finds the low-dimensional embedding of the data[11]. The Laplacian eigenmaps is a computationally effective and map nearby input patterns to nearby outputs by computing the low-dimensional representation of a high-dimensional data set that most faithfully preserves proximity relations and it has a natural connection with clustering[2]. PCA is a popular linear technique used for feature extraction or dimensionality reduction. Given a set of data with n dimensions, PCA maps the data linearly to find a subspace in lower-dimensional space so that variance of the data is maximized. It does so by calculating the eigenvectors from the covariance matrix. The principal components (eigenvectors that correspond to the largest eigenvalues) are used to recreate a substantial portion of the original data's variance[12]. t-SNE is a non-linear dimensionality reduction technique. t-SNE is not used for cluster analysis or outlier detection since it does not preserve the data's distances or densities. But, it is particularly well suited for the visualization of high-dimensional datasets and extensively applied in image processing, Natural language processing, genomic data, and speech processing [33].

2.3.4 Clustering

Performing clustering is one of the critical tasks in single-cell analysis. Clusters are formed by grouping cells based on their similarity of the gene expression profiles. Distance metrics are used to describe expression profile similarity, which employs dimensionality-reduced representations as data as input. We used popular clustering technique k-means. k-means is iterative clustering algorithm groups the data into n separate groups by minimizing the phenomenon within-cluster dispersion. The number of clusters $\mathbf{C} = \{\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_n\}$ to be formed from the data needs to be specified as an input to the algorithm.

$$SSE = \sum_{i=1}^{k} \min_{\mu \in C} (|\mathbf{x}_i - \mathbf{x}_j|)^2 \tag{6}$$

k-means algorithm works in three key steps. The first step is to choose the initial centroids and the simple method is to choose k samples from the dataset $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$. Then, each point in the dataset is allocated to its nearest centroid. The next step involves taking the mean value of all of the samples allocated to each previous centroid and creating new centroids. The algorithm calculates the difference between the old and new centroids, then repeats the last two steps until the value falls below a certain threshold. In other words, it keeps repeating until the centroids are converged. The points in the data choose centroids with a high degree of cluster compactness or a minimum sum of squared error (SSE) as shown in (6) where n is the number of samples in the data, C is the cluster, μ is the mean of the samples, and x is the corresponding sample.

2.3.5 Cluster Annotation

Gene Set Enrichment Analysis (GSEA) [31] is a computational tool that determines whether a predefined set of genes shows a statistically significant level of expression in a specific cell type, biological process, cellular component, molecular function, or biological pathway. The GSEA uses MSigDB, the Molecular Signature Database, to provides different gene sets for the analysis with the gene set enrichment analysis. To annotate the cell clusters, we first extracted the top 20 differentially expressed genes as markers in each cluster per dataset. Then, we found the corresponding cell types of each group of marker genes in each cluster. Gene ontology (GO) analysis is also used as part of enrichment analysis.

2.3.6 Parameter Optimization

With the aim of preserving locality, the number of nearest neighbors (t-NN) to construct the neighborhood graph is a crucial parameter in manifold learning techniques.

Another critical step in any clustering algorithm is determining the number of clusters, k. We used the nearest neighbor check and validity indices check, which runs through different t and a distinct number of clusters to find the best dimensionality reduction and clustering parameters. We further systematically evaluated more appropriate parameters for MLLE after finding the best t. The nearest neighbors are examined between the range of 8 and 26. The number of clusters k for each value of t is also assessed, where k ranges from 4 to 15, and the validity of indices are calculated for each cluster. We select a combination of t and the number of clusters with the highest number of clustering scores considering all the three validity of indices explained in the performance evaluation section.

2.3.7 Performance Evaluation

Generally speaking, the best clustering is the one that maintains high intra-cluster distance and gives the most compact clusters. In this work, we use the Silhouette coefficient [26], an evaluation metric that measures either the mean distance between a sample point and all other points in the same cluster or all other points in the next nearest neighbor cluster. Consider a set of clusters $\mathbf{C} = \{\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_k\}$, output by a clustering algorithm, k-means in our case. The Silhouette coefficient, SH, for the i^{th} sample point in cluster $\mathbf{C}_j$, where $j = 1, ..., k$, can be defined as follows:

$$SH(\mathbf{x}_i) = \frac{b(\mathbf{x}_i) - a(\mathbf{x}_i)}{max(a(\mathbf{x}_i), b(\mathbf{x}_i))} \, , \tag{7}$$

where a is the mean distance between point $\mathbf{x}_i$ and all other points inside the cluster (intra-cluster distance) and b is the minimum mean value of the distance between a sample point $\mathbf{x}_i$ and the nearest neighbor cluster, and are calculated as:

$$\begin{aligned} a(\mathbf{x}_i) &= \frac{1}{|\mathbf{C}_k| - 1} \sum_{\mathbf{x}_j \in \mathbf{C}_k, i \neq j} d(\mathbf{x}_i, \mathbf{x}_j) \\ b(\mathbf{x}_i) &= \min_{k \neq i} \frac{1}{|\mathbf{C}_k|} \sum_{j=1}^{k} d(\mathbf{x}_i, \mathbf{x}_j)\,. \end{aligned} \tag{8}$$

We also used Calinski-Harabasz (CH) and Davies-Bouldin (DB) validity of indices to assess the clustering performance. Calinski-Harabasz score [3], is a score used to evaluate the model where a higher score tells better-defined clusters. CH score is the ratio of the sum of between-clusters dispersion and of inter-cluster dispersion for all clusters that is as follows:

$$CH = \frac{tr(\mathbf{S}_B)}{tr(\mathbf{S}_W)} \times \frac{n-k}{k-1} \tag{9}$$

in which n is size of input samples, $tr(\mathbf{S}_B)$ is the trace of the between-group dispersion matrix and $tr(\mathbf{S}_W)$ is the within-cluster dispersion.

Davies-Bouldin (DB) index [5] is another validity index defined as the average of the similarity measure of each cluster. DB is computed as follows:

$$DB = \frac{1}{k} \sum_{i=1}^{k} max_{i \neq j} s_{ij}\,, \tag{10}$$

where s_{ij} is the ratio between within-cluster distances and between cluster distances, and is calculated as $s_{ij} = \frac{w_i + w_j}{d_{ij}}$. The smaller DB value the better clustering, and as such, we aim to minimize Equation (10). Here, d_{ij} is the Euclidean distance between cluster centroids μ_i and μ_j, and w_i is the within-cluster distance of cluster $\mathbf{C}_k$.

Overall, we used the Silhouette score to evaluate the clustering performance, whereas CH and DB indices were used to verify and find the optimal parameters, namely the best number of clusters.

2.4 Results and Discussion

We developed a well-constructed pipeline that can be applied to single-cell data to discover individual cell types. Considering dimensionality reduction and clustering as two significant steps in the pipeline, we conducted many experiments on different dimensionality reduction techniques and explored many ways of untangling the data in two and three dimensions. We found optimum parameters for both dimensionality reduction and clustering to achieve the best clustering results. To demonstrate the applicability of our pipeline, we tested it on thirteen datasets of different sizes. We evaluated our method in terms of both computationally and biologically perspectives to achieve the meaningful separation of cell types.

2.4.1 Clustering and Cell Type Discovery

To achieve the best results, we experimented with all possible combinations of parameters as discussed in the Material and Methods section. As a result, the best parameters chosen for each dataset are depicted in Table 2.4.1. In a few datasets, to achieve the best clustering score in the proposed approach, the data is reduced to lower dimensions such as 5, 6, and 7. Then, the data is reduced to three dimensions to visualize and obtain better results. When applying MLLE, a neighborhood graph is created by connecting points that are close to each other. Different measures are used for this purpose, including number of neighbors, distance from each point to its neighbors, and others. A common measure to determine the sparsity of the neighbor graph is the tolerance factor, which makes the graph sparser or denser. In this regard, we tested different tolerance values on each dataset and selected those values that yielded the best validity index scores. The results of k-means clustering combined with each dimensionality reduction method using the best parameters are listed in Table 2.4.2. The last column shows the result after applying ICA on the result of clustering combined with MLLE. The clustering score ranges from 0 to 1. A score close to 1 represents good quality clustering, with 1 being the best, while a score near zero indicates that the clusters are not well defined.

Table 2.4.1: Parameters used for experiments. These are generated considering both dimensionality reduction and clustering together.

Dataset name	No. of Neighbors	No. of Dimensions	Tolerance	No. of Clusters
Baron_human1	10 23	6 3	1e-12 1e-10	14
Baron_human2	8	3	1e-12	14
Baron_human3	16 8	7 3	1e-12 1e-8	14
Baron_human4	9 22	6 3	1e-12 1e-12	14
Baron_mouse1	17	3	1e-12	13
Baron_mouse2	11 20	6 3	1e-12 1e-8	13
Muraro	10 11	5 3	1e-3 1e-7	6
Segerstolpe	10 9	5 3	1e-3 1e-8	6
Xin	15 25	6 3	1e-12 1e-3	6
Wang	8	3	1e-12	6
H1299_scRNAseq	11	3	1e-8	7
Calu3_scRNAseq	12 11	7 3	1e-3 1e-5	7
PBMC	8 25	5 3	1e-12 1e-12	8

Table 2.4.2: Silhoutte scores comparison of proposed method with other dimensionality reduction techniques.

Dataset name	t-SNE	PCA	Isomap	SLLE	Eigenmaps	MLLE	MLLE+ICA
Baron_human1	0.244	0.364	0.498	0.524	0.839	**0.908**	0.904
Baron_human2	0.231	0.428	0.543	0.614	0.823	**0.906**	0.905
Baron_human3	0.243	0.377	0.522	0.467	0.826	**0.990**	0.976
Baron_human4	0.239	0.424	0.614	0.538	0.896	0.910	**0.912**
Baron_mouse1	0.231	0.400	0.422	0.448	0.472	0.881	**0.917**
Baron_mouse2	0.221	0.414	0.530	0.684	0.779	0.941	**0.943**
Muraro	0.258	0.494	0.532	0.738	0.913	0.933	**0.944**
Segerstolpe	0.231	0.410	0.399	0.400	0.537	**0.960**	0.956
Xin	0.242	0.445	0.481	0.494	0.751	**0.899**	0.888
Wang	0.230	0484	0.442	0.745	0.608	0.993	**0.996**
H1299_scRNAseq	0.245	0.269	0.701	0.683	0.782	0.938	**0.943**
Calu3_scRNAseq	0.361	0.232	0.494	0.452	0.798	0.889	**0.924**
PBMC	0.244	0.401	0.622	0.621	0.632	0.867	**0.876**

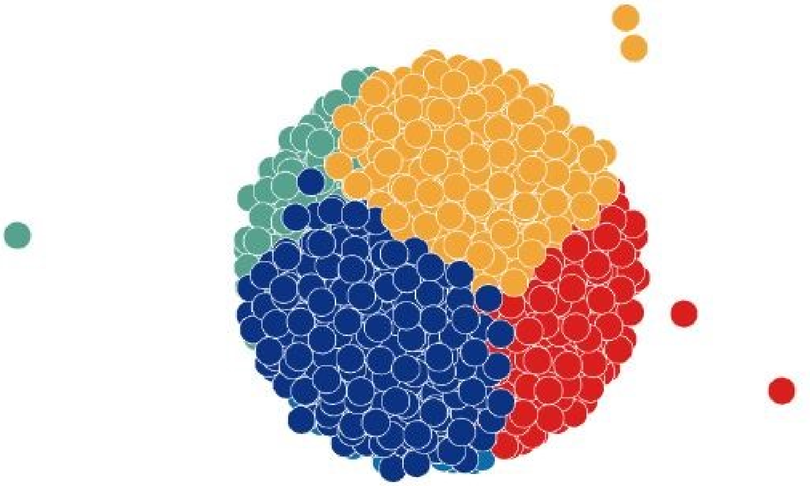

Fig. 2.4.1: Visualization of t-SNE on Muraro dataset

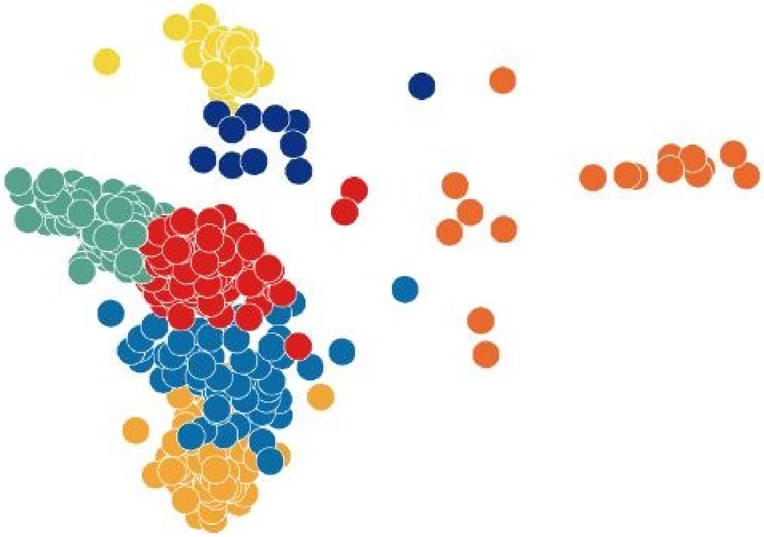

Fig. 2.4.2: Visualization of PCA on Wang dataset

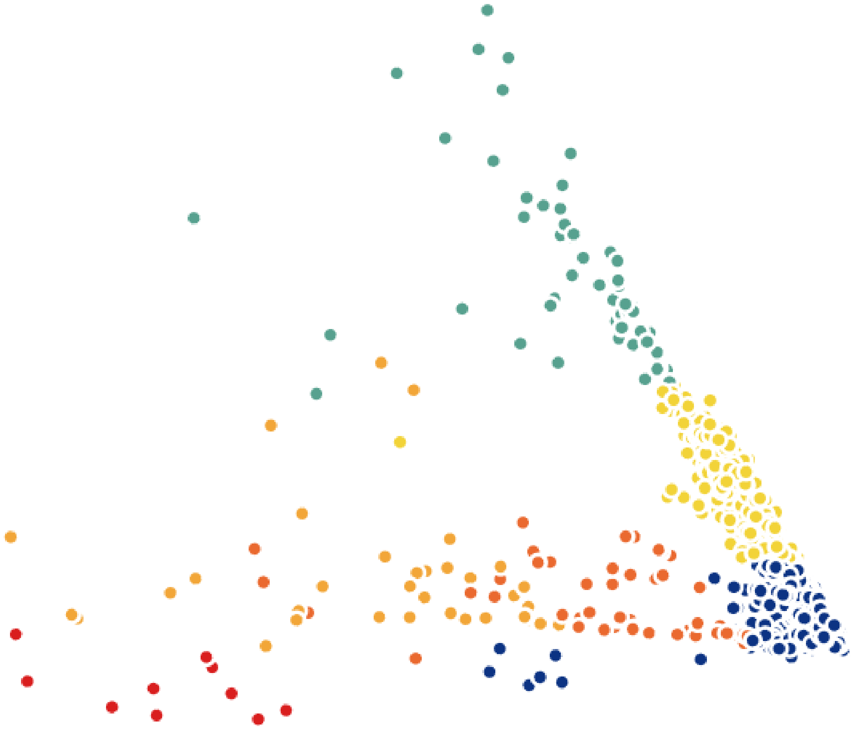

Fig. 2.4.3: Visualization of Laplacian eigenmaps on H1299_scRNAseq; outliers have been removed to enhance visualization.

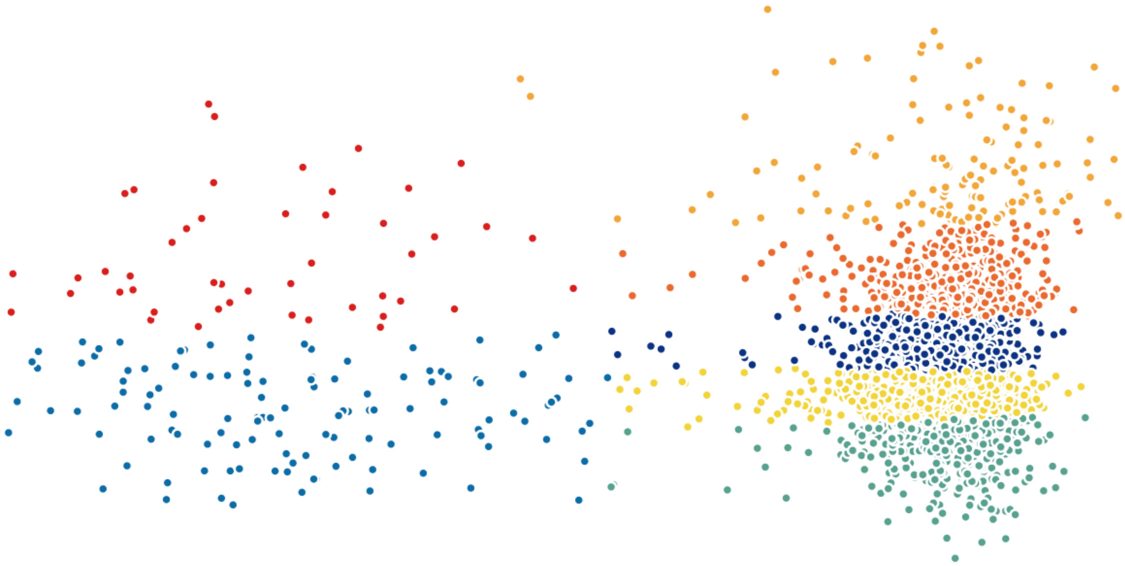

Fig. 2.4.4: Two-dimensional ICA projection of cells colored by k-means clustering applied on high-dimensional original data (H1299_scRNAseq).

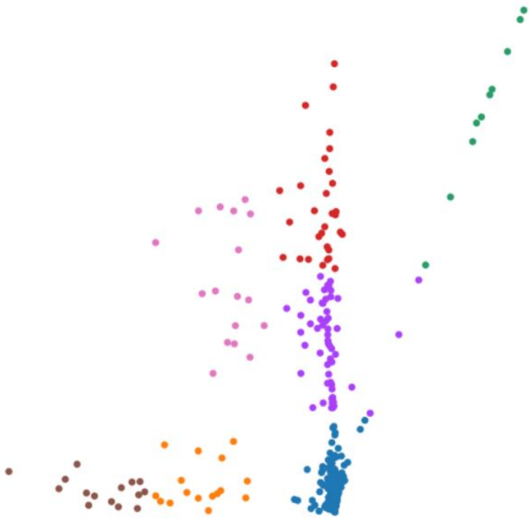

Fig. 2.4.5: Two-dimensional ICA projection of cells colored by k-means clustering applied to the three-dimensional points output by MLLE on the H1299_scRNAseq dataset.

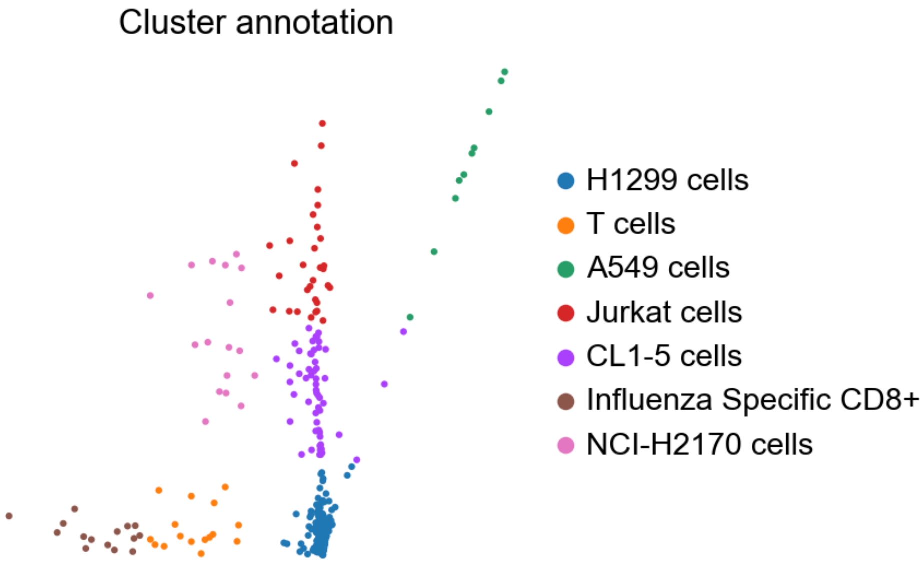

Fig. 2.4.6: Cluster annotation for H1299_scRNAseq.

When trying widely-used techniques such as t-SNE and PCA, we noticed that both methods were not as efficient in separating the data into well-defined clusters. To show the clustering results graphically, we visualize the result of PCA and t-SNE for Wang and Muraro datasets, respectively, in Figures 2.4.1 and 2.4.2. On the other hand, the results of Isomap and Standard LLE show slightly better performance comparatively. Moreover, Laplacian Eigenmaps performed better than these two methods, though they could not accomplish competitive clustering. As a good example, we visualize samples from H1299_scRNAseq using Laplacian Eigenmaps (Fig.2.4.3) in which different clusters are overlapping. Finally, we investigated MLLE and found the most insightful cluster separation in most of the datasets. This outcome demonstrates the power of MLLE in exploring the data's dense and complex relations, creating better lower embeddings. We performed an additional dimensionality reduction step that uses ICA to enhance the visualization of clusters. The last column of Table 2.4.2 represent that MLLE combined with ICA improves the overall results except for some datasets that we can not see much difference; very negligible difference of 0.004 (Baron_human1), 0.001 (Baron_human2), 0.014 (Baron_human3), 0.004 (Segerstolpe), and 0.011 (Xin) can ignore them. To achieve a better view of the

impact of ICA on the MLLE transformation, we show a visual comparison of clusters in Figures 2.4.4 and 2.4.5. Two-dimensional ICA projection of the cells applied to the three-dimensional MLLE data shows the best visualization and clustering scores (Fig. 2.4.5). When applied alone, ICA performed very poorly with significantly inseparable clusters (Fig. 2.4.4). This is because ICA is limited to linear transformations.
On the other hand, manifold learning techniques consider data locally. As such, it can reveal complex relationships among the data points in higher-dimensional spaces. We instead applied ICA on the lower-dimensional data because we observed well-marked "lines" or "axes" in the three-dimensional data, which led us to think that we could apply ICA to learn the linearly independent components, not necessarily orthogonal. Applying ICA reveals some hidden, complex relationships among the cells in the clusters, which are not noticeable in three dimensions.

2.4.2 Biological Assessment

To validate the obtained clusters, we first identified the top 20 genes in each cluster based on the Wilcoxon test. Starting from these top 20 genes, we retrieved a subset of genes from the largest number of overlapping genes across the different clusters. Marker genes are up- or down-regulated in different individual cells, pathways or GO terms. We used GSEA and ToppCluster multi-gene list functional enrichment analysis online tools to identify GO terms and pathways associated with the top 20 gene lists extracted from each cluster. Pathways were extracted from the MSigDB C2 BIOCARTA (V7.3) database [21]. Cytoscape [30] was used to visualize the networks.We decreased the minimum number of genes present in annotations to achieve a better visualization.

As presented in Table 2.4.5, some of the pancreatic cell types are found for pancreas datasets such as the Baron human dataset within well-defined gene sets in MSigDB namely 'MURARO PANCREAS ALPHA CELL', 'MURARO PANCREAS ENDOTHELIAL CELL', 'MURARO PANCREAS MESENCHYMAL STROMAL CELL', 'MURARO PANCREAS DUCTAL CELL', and 'MURARO PANCREAS ACINAR CELL'. Other cell types including CD34, Jurkat, and macrophage are cell

Table 2.4.3: Identified cell types for H1299_scRNAseq.

Cell Types	Cluster Number
H1299 cells	0
T cells	1
A549 cells	2
Jurkat cells	3
CL1-5 cells	4
Influenza-specific CD8+	5
NCI-H2170 cells	6

subtypes of T-Cells. HB2 is also a cell line originated by epithelial cells. Regarding H1299_scRNAseq and Calu3_scRNAseq datasets, Tables 2.4.3 and 2.4.4 list associated cell types mostly involved in the immune system. It is well-known that one of the main SARS-CoV-2 targets is the immune system function. We observed co-expressed gene sets down- or up-regulated in the lung and immune systems specific cell (sub)types. T-cell is a type of immune cell that is found in blood. Jurkat cells are a line of human T cells that are used to study the expression of various chemokine receptors susceptible to viral entry, particularly HIV. CD8+ T cells are found on the surface of immune cells and are key cells in response to viral infection [4]. Moreover, H1299 cells, NCI-H2170 cells, A549 cells, and CL1-5 cells are human lung associated cell lines. These findings show the effectiveness of the proposed method to identify associated cell types using cell type specific marker genes. A projection of the identified cells in H1299_scRNAseq colored by clusters is shown in Fig. 2.4.6.

Additionally, visualization of GO terms and pathways associated with the corresponding marker genes are depicted in Figs. 2.4.7 and 2.4.8, respectively. For each cluster, we identified a set of biological process or pathway terms. Each edge in the plot shows a link between a cluster and a term that is significantly associated with the 20 top gene list in that cluster.By observing Fig. 2.4.8, some significant pathways

Table 2.4.4: Identified cell types for Calu3_scRNAseq.

Cell Types	Cluster
H1299 cells	0
293 cells (embryonic kidney)	1
MCF7 cells	2
ANBL-6 cell	3
T-ALL	4
H460 cells	5
H1975 cells	6

Table 2.4.5: Identified cell types for Baron_human1 dataset.

Cell Types	Cluster
Alpha	0
CD34	1
Mesenchyme stem cells	2
Jurkat cells (T lymphocyte)	3
Endothelial	4
Mesenchyme stromal cells	5
Ductal	6
Endothelial	7
Acinar	8
Myeloid cells	9
Intestine cells	10
Macrophage	11
HB2 cells	12
T-cells	13

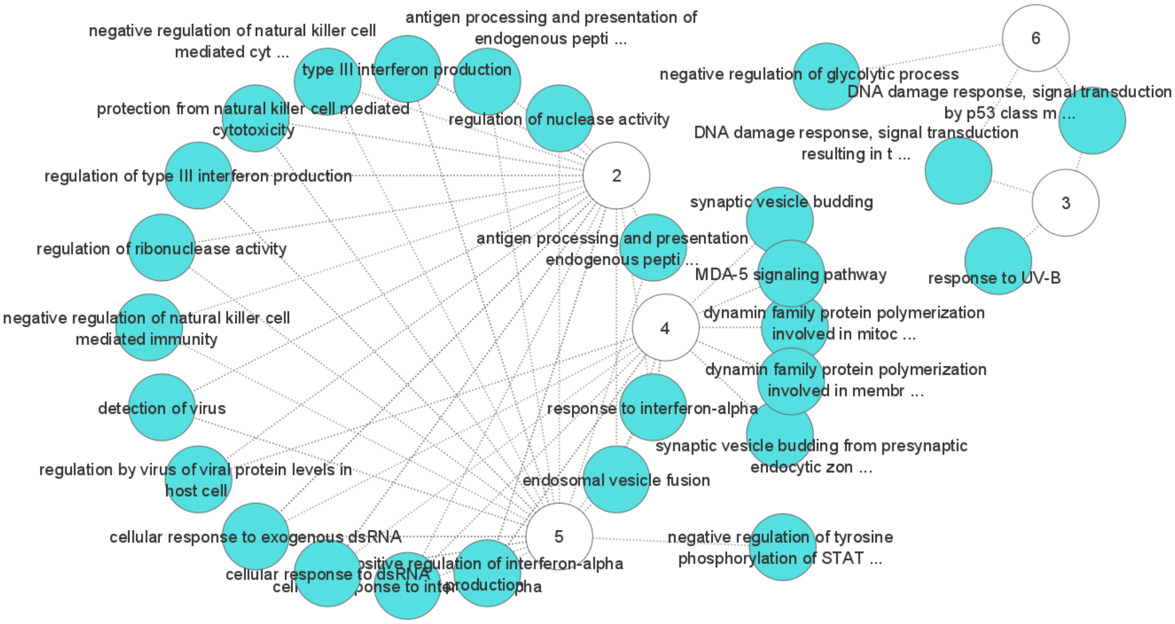
negative regulation of natural killer cell mediated cyt ...
antigen processing and presentation of endogenous pepti ...
type III interferon production
6
negative regulation of glycolytic process
DNA damage response, signal transduction by p53 class m ...
protection from natural killer cell mediated cytotoxicity
regulation of nuclease activity
DNA damage response, signal transduction resulting in t ...
regulation of type III interferon production
2
3
synaptic vesicle budding
regulation of ribonuclease activity
antigen processing and presentation endogenous pepti ...
MDA-5 signaling pathway
response to UV-B
negative regulation of natural killer cell mediated immunity
4
dynamin family protein polymerization involved in mitoc ...
dynamin family protein polymerization involved in membr ...
detection of virus
response to interferon-alpha
synaptic vesicle budding from presynaptic endocytic zon ...
regulation by virus of viral protein levels in host cell
endosomal vesicle fusion
cellular response to exogenous dsRNA
5
negative regulation of tyrosine phosphorylation of STAT ...
cellular response to dsRNA
positive regulation of interferon-alpha production

Fig. 2.4.7: A set of biological process that are enriched by marker genes in H1299_scRNAseq dataset. The numbers show the clusters and edges shows the link between a cluster and a biological process term.

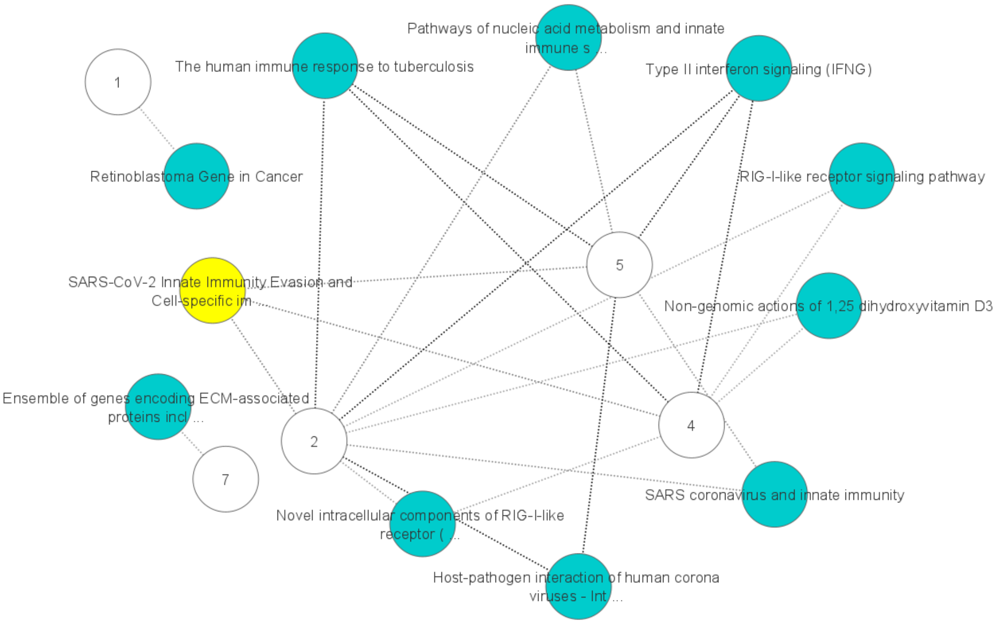
Pathways of nucleic acid metabolism and innate immune s ...
1
The human immune response to tuberculosis
Type II interferon signaling (IFNG)
Retinoblastoma Gene in Cancer
RIG-I-like receptor signaling pathway
5
SARS-CoV-2 Innate Immunity Evasion and Cell-specific im ...
Non-genomic actions of 1,25 dihydroxyvitamin D3
Ensemble of genes encoding ECM-associated proteins incl ...
4
2
7
SARS coronavirus and innate immunity
Novel intracellular components of RIG-I-like receptor (...
Host-pathogen interaction of human corona viruses - Int ...

Fig. 2.4.8: Pathway that are enriched by marker genes in H1299_scRNAseq dataset. The numbers show the clusters and edges shows the link between a cluster and a pathway. Node that is highlighted yellow show the SARS-CoV-2 cell-specific pathway. Most of the other green nodes reveal the shared and cluster-specific functional pathways in the immune system.

are found to be enriched in immunity functions, and signaling identified, including SARS-CoV-2 innate Immunity Evasion, Host-pathogen interaction of human corona viruses, SARS coronavirus and innate immunity, Type II interferon signaling (IFNG), and the human immune response to tuberculosis. Also, the gene set enrichment of Fig. 2.4.7 shows that most biological processes are associated with immunity functions, including response to interferon-alpha, protection from a natural killer cell, type III interferon production, regulation by virus of viral protein levels in a host cell, and detection of virus, among others. In addition, we obtained a list of overlapping marker genes that are involved in Herpes simplex virus 1 (HSV-1) infection and the Influenza A pathway. These findings suggest potential markers for subsequent medical treatment or drug discovery by comparing to similar diseases in terms of functionality. Moreover, although numerous findings suggest potential links between HSV-1 and Alzheimer's disease, a causal relationship has not been demonstrated yet [6].

The outcomes of this work can be summarized as follows. Performing ICA on transformed data after applying manifold learning techniques provides improved the clustering output and meaningful organization of cell clusters. Moreover, modified LLE combined with k-means leads to an enhanced view of the data and the corresponding clusters by "untangling" the complex, hidden relationship in a higher-dimensional space. Such non-linear dimensionality reduction methods have been shown to be very powerful as they preserve the locality of the data from higher to lower dimensions. Evaluating the incidence of ICA as a visualization scheme and further reduction step shows better clustering and enhanced visualization simultaneously. This trend leads to a research avenue that involves a combination of non-linear manifold learning techniques followed by linear methods, which has shown to be more powerful than conventional methods such as PCA or ICA applied alone.

References

[1] Maayan Baron et al. “A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure”. In: *Cell systems* 3.4 (2016), pp. 346–360.

[2] Mikhail Belkin and Partha Niyogi. “Laplacian eigenmaps for dimensionality reduction and data representation”. In: *Neural computation* 15.6 (2003), pp. 1373–1396.

[3] Tadeusz Caliński and Jerzy Harabasz. “A dendrite method for cluster analysis”. In: *Communications in Statistics-theory and Methods* 3.1 (1974), pp. 1–27.

[4] Eddie Cano-Gamez et al. “Single-cell transcriptomics identifies an effectorness gradient shaping the response of CD4+ T cells to cytokines”. In: *Nature communications* 11.1 (2020), pp. 1–15.

[5] David L Davies and Donald W Bouldin. “A cluster separation measure”. In: *IEEE transactions on pattern analysis and machine intelligence* 2 (1979), pp. 224–227.

[6] Giovanna De Chiara et al. “Recurrent herpes simplex virus-1 infection induces hallmarks of neurodegeneration and cognitive deficits in mice”. In: *PLoS pathogens* 15.3 (2019), e1007617.

[7] Chuan Dong et al. “Comprehensive review of the identification of essential genes using computational methods: focusing on feature implementation and assessment”. In: *Briefings in bioinformatics* 21.1 (2020), pp. 171–181.

[8] Albert Einstein, Boris Podolsky, and Nathan Rosen. “Can quantum-mechanical description of physical reality be considered complete”. In: *Physical review* 47.10 (1935), p. 777.

[9] Chao Feng et al. “Dimension reduction and clustering models for single-cell rna sequencing data: A comparative study”. In: *International journal of molecular sciences* 21.6 (2020), p. 2181.

[10] 10X Genomics. *Single Cell Gene Expression Dataset by Cell Ranger 1.1.0*. May 2016.

[11] Ali Ghodsi. “Dimensionality reduction a short tutorial”. In: *Department of Statistics and Actuarial Science, Univ. of Waterloo, Ontario, Canada* 37.38 (2006), p. 2006.

[12] Ali Ghodsi. “Dimensionality reduction a short tutorial”. In: *Department of Statistics and Actuarial Science, Univ. of Waterloo, Ontario, Canada* 37.38 (2006), p. 2006.

[13] Dominic Grün et al. “Single-cell messenger RNA sequencing reveals rare intestinal cell types”. In: *Nature* 525.7568 (2015), pp. 251–255.

[14] Manuel Guerrero et al. “Adaptive community detection in complex networks using genetic algorithms”. In: *Neurocomputing* 266 (2017), pp. 101–113.

[15] Byungjin Hwang, Ji Hyun Lee, and Duhee Bang. “Single-cell RNA sequencing technologies and bioinformatics pipelines”. In: *Experimental & molecular medicine* 50.8 (2018), pp. 1–14.

[16] Aapo Hyvärinen. “Independent component analysis: recent advances”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 371.1984 (2013), p. 20110534.

[17] Aapo Hyvärinen and Erkki Oja. “Independent component analysis: algorithms and applications”. In: *Neural networks* 13.4-5 (2000), pp. 411–430.

[18] Tomislav Ilicic et al. “Classification of low quality cells from single-cell RNA-seq data”. In: *Genome biology* 17.1 (2016), pp. 1–15.

[19] Saiful Islam et al. “Quantitative single-cell RNA-seq with unique molecular identifiers”. In: *Nature methods* 11.2 (2014), p. 163.

[20] Vladimir Yu Kiselev, Tallulah S Andrews, and Martin Hemberg. “Challenges in unsupervised clustering of single-cell RNA-seq data”. In: *Nature Reviews Genetics* 20.5 (2019), pp. 273–282.

[21] Arthur Liberzon et al. “Molecular signatures database (MSigDB) 3.0”. In: *Bioinformatics* 27.12 (2011), pp. 1739–1740.

[22] Malte D Luecken and Fabian J Theis. “Current best practices in single-cell RNA-seq analysis: a tutorial”. In: *Molecular systems biology* 15.6 (2019), e8746.

[23] Leland McInnes, John Healy, and James Melville. “Umap: Uniform manifold approximation and projection for dimension reduction”. In: *arXiv preprint arXiv:1802.03426* (2018).

[24] Mauro J Muraro et al. “A single-cell transcriptome atlas of the human pancreas”. In: *Cell systems* 3.4 (2016), pp. 385–394.

[25] Xiaojie Qiu et al. “Single-cell mRNA quantification and differential analysis with Census”. In: *Nature methods* 14.3 (2017), pp. 309–315.

[26] Peter J Rousseeuw. “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis”. In: *Journal of computational and applied mathematics* 20 (1987), pp. 53–65.

[27] Rickard Sandberg. “Entering the era of single-cell transcriptomics in biology and medicine”. In: *Nature methods* 11.1 (2014), pp. 22–24.

[28] Lawrence K Saul and Sam T Roweis. “An introduction to locally linear embedding”. In: *unpublished. Available at: http://www. cs. toronto. edu/˜ roweis/lle/publications. html* (2000).

[29] Åsa Segerstolpe et al. “Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes”. In: *Cell metabolism* 24.4 (2016), pp. 593–607.

[30] Paul Shannon et al. “Cytoscape: a software environment for integrated models of biomolecular interaction networks”. In: *Genome research* 13.11 (2003), pp. 2498–2504.

[31] Aravind Subramanian et al. "GSEA-P: a desktop application for Gene Set Enrichment Analysis". In: *Bioinformatics* (). DOI: `10.1093/bioinformatics/btm369`. eprint: `https://academic.oup.com/bioinformatics/article-pdf/23/23/3251/16860704/btm369.pdf`. URL: `https://doi.org/10.1093/bioinformatics/btm369`.

[32] Tatiana Tatusova et al. "NCBI prokaryotic genome annotation pipeline". In: *Nucleic acids research* 44.14 (2016), pp. 6614–6624.

[33] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11 (2008).

[34] Jianzhong Wang. "Laplacian Eigenmaps". In: *Geometric Structure of High-Dimensional Data and Dimensionality Reduction*. Springer, 2012, pp. 235–247.

[35] Yue J Wang et al. "Single-cell transcriptomics of the human endocrine pancreas". In: *Diabetes* 65.10 (2016), pp. 3028–3038.

[36] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. "SCANPY: large-scale single-cell gene expression data analysis". In: *Genome biology* 19.1 (2018), pp. 1–5.

[37] Emanuel Wyler et al. "Bulk and single-cell gene expression profiling of SARS-CoV-2 infected human cell lines identifies molecular targets for therapeutic intervention". In: *BioRxiv* (2020).

[38] Yurong Xin et al. "RNA sequencing of single human islet cells reveals type 2 diabetes genes". In: *Cell metabolism* 24.4 (2016), pp. 608–615.

[39] Christopher Yau et al. "pcaReduce: hierarchical clustering of single cell transcriptional profiles". In: *BMC bioinformatics* 17.1 (2016), pp. 1–11.

[40] Amit Zeisel et al. "Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq". In: *Science* 347.6226 (2015), pp. 1138–1142.

CHAPTER 3

Conclusion and Future Work

3.1 Conclusion

We have proposed a method that focuses on identifying different cell types using powerful manifold learning for dimensionality reduction combined with independent component analysis and clustering techniques on scRNA-seq data. This pipeline accommodates all the requirements according to standard scRNA-seq protocols to perform downstream analysis such as quality metrics evaluation, dimensionality reduction, and clustering. We have conducted extensive experiments to find the optimized parameters for dimensionality reduction and clustering by finding the number of nearest neighbors and the number of clusters, respectively. Efficient nonlinear dimensionality reduction and manifold learning techniques significantly improve the clustering results, and the linear ICA method enhances visualization in a reduced space. Using multiple benchmark datasets shows our proposed method's general accuracy and turns out to be a promising approach for discovering cell types. Performing gene set enrichment analysis to annotate a set of highly-variable genes obtained from each cluster reveals biomarker genes involved in different gene ontology terms.

Overall, we delineated a pipeline to highlight the power of a combination of linear methods such as ICA and manifold learning techniques to find cell types and validated it using various benchmark datasets.

3.1.1 Contributions

The main contributions of this thesis can be summarized as follows:

- Proposed a new validated pipeline used to identify cell types in single-cell data.
- Proposed a new combination and a hybrid model for dimensionality reduction using linear and non-linear methods, which is powerful in exploring the data from higher to lower dimensions.
- Proposed a mechanism to choose the optimized combination of number of nearest neighbors and the number of clusters k for clustering and dimensionality reduction using validity of indices.
- Developed a software tool for identifying cell types in single-cell data and a GitHub project.

3.2 Future Work

Currently, the proposed method identifies cell types by performing efficient dimensionality reduction and clustering. This is achieved through manual investigation of top maker genes in biological databases, which is tedious and might not obtain accurate results sometimes. This work can be extended in several ways, some of which being listed as follows:

- Automatic identification of cell types reduces a great deal of time and effort because, in the manual investigation, the user has to take all sets of genes and query in databases to get the results, and evaluating those results is tedious.
- More genes can be selected for the downstream analysis, which contributes more information using different strategies for selecting genes or employing feature selection methods.
- Autoencoders is one of the neural networks that can be employed to learn the representation of the data in lower dimensions. It can be used for further analysis, which might create more clear representations of the data and improve clustering results.

- t-SNE performs well in visualizing the data, and future research can be performed to improve visualization and clustering with hybrid models considering other dimensionality reduction techniques such as Laplacian eigenmaps and LLE.

- Self Organizing Maps (SOMs) can be applied for both dimensionality reduction and clustering, which creates more meaningful neighborhood relations even in lower-dimensions.

- As a further analysis in the future, one can perform other epigenetics challenges and emerging directions in single-cell analysis, such as trajectory and pathway analyses.

APPENDIX A

Marker Genes

A marker gene is a DNA sequence that has been identified on a chromosome. Marker genes also help in the identification of the gene responsible for an inherited disease. On a chromosome, DNA segments similar to each other are more likely to be inherited. Marker genes are used to trace the inheritance of a nearby gene that has yet to be discovered but is considered close. The marker genes obtained from our research are found in few diseases. These marker genes shown in table A.0.1 are helpful in studying the diseases. Our marker genes list is overlapped in Herpes simplex virus 1 (HSV-1) infection and Influenza A pathway.

Table A.0.1: Marker genes found in similar diseases.

Disease	**Marker Genes**
Influenza	RSAD2, IFIH1, MX1, STAT1, MX2, IRF7, TNFSF10, OAS1, DDX58, NFKBIA, OAS2, CXCL10, EIF2AK2, PML, ICAM1, CXCL8, OAS3, STAT2
Herpes Simplex Virus 1	IFIH1,HLA-B,STAT1,IRF7 TAP1,OAS1, DDX58,NFKBIA, OAS2,STAT2,EIF2AK2, SP100, PML,HLA-E,B2M,OAS3,HLA-F

These results can be used for subsequent medical treatment or drug discovery by finding similar diseases in functionality. Moreover, although numerous findings suggested potential links between HSV-1 and Alzheimer's disease (AD), a causal relationship has not been demonstrated yet.

VITA AUCTORIS

NAME:	Saiteja Danda
PLACE OF BIRTH:	Tirumani, Karnataka, India
EDUCATION:	Bachelors in Computer Science and Engineering, Visvesvaraya Technological University, Belgaum, Karnataka, India, 2017 M.Sc. Computer Science, University of Windsor, Windsor, Ontario, Canada, 2020