國立陽明交通大學

生物醫學資訊研究所

碩士論文

Institute of Biomedical Informatics

National Yang Ming Chiao Tung University

Master Thesis

整合單細胞 RNA 定序資料
以模擬去除雜訊的傳統 RNA 定序資料

Ensemble of scRNA-seq data
to simulate denoised bulk RNA-seq data

研究生：鐘偉哲 (CHUNG, WEI-CHE)

指導教授：黃宣誠 (HUANG, HSUAN-CHENG)

中華民國一一○年七月

July 2021

整合單細胞 RNA 定序資料
以模擬去除雜訊的傳統 RNA 定序資料
Ensemble of scRNA-seq data
to simulate denoised bulk RNA-seq data

研 究 生：鐘偉哲　　　　Student：Wei-Che Chung

指導教授：黃宣誠 博士　　Advisor：Dr. Hsuan-Cheng Huang

國立陽明交通大學

生物醫學資訊研究所

碩士論文

A Thesis Submitted to
Institute of Biomedical Informatics
College of Medicine and
College of Life Sciences
National Yang Ming Chiao Tung University
in Partial Fulfillment of the Requirements
for the Degree of
Master of Science
in Biomedical Informatics

July 2021
Taiwan, Republic of China

中華民國 一一○年七月

摘要

本研究嘗試通過演算法將單細胞 RNA 定序資料模擬成傳統 RNA 定序資料。單細胞 RNA 定序在 2009 年被首次發表，時至今日由於定序成本下降，大量的單細胞 RNA 定序資料進入研究者們的視野中，單細胞 RNA 定序的出現讓人類可以研究新的生物學問題，例如：細胞類型鑑定和癌細胞等等異質性細胞的研究，其原因來自單細胞 RNA 定序所產生的基因表現量矩陣讓我們可以觀察特定細胞的變化。由於現階段出現了大量的單細胞 RNA 定序資料集，許多在過去適用於傳統 RNA 定序的方法出現了不適用的情況，例如：基因共表現分析方法(WGCNA)。因為傳統 RNA 定序的基因表現量矩陣相對單細胞 RNA 定序的基因表現量矩陣來說，傳統 RNA 定序的基因表現量矩陣具有雜訊較小、空值較少等特色，單細胞 RNA 定序則因為定序實驗步驟較多，導致雜訊較大，同時定序的深度較淺，所以空值較多；由於兩者 RNA 定序方法基本上是一樣的，所以本研究嘗試使用演算法將單細胞 RNA 定序基因表現量矩陣的雜訊減少以及減少空值，使其達到模擬傳統 RNA 定序基因表現量矩陣的目的。

本研究的方法可以為每個單細胞 RNA 定序樣本生成模擬數據，每個模擬樣本都具有與傳統 RNA 定序相同的屬性，包括雜訊、空值的減少和樣本間相關性、低表現量基因之數值的增加。

本研究的方案可以用於多種研究情景。例如，拿單細胞定序資料和既存的傳統 RNA 定序資料做比較。從數據分析的方面來說，我們可以產生大量且低雜訊的類傳統 RNA 定序樣本，比如使用癌細胞的單細胞定序資料模擬成單純的低雜訊癌組織定序樣本。可以避免癌組織的傳統 RNA 定序樣本時常混雜不同細胞類型的問題。

Abstract

This research attempts to develop an algorithm to simulate traditional bulk RNA sequencing (bulk RNA-seq) data by single-cell RNA sequencing (scRNA-seq) data. scRNA-seq was first published in 2009. Nowadays, due to the decline in sequencing costs, a large amount of scRNA-seq data have been generated and provided to researchers. The emergence of scRNA-seq allows researchers to study new biology issues. For example, identification of cell types and the study of heterogeneous cells, such as cancer cells, are due to scRNA-seq. This allows us to observe expression in specific cells. Currently, there are several traditional bulk RNA-seq datasets and many methods suitable for traditional bulk RNA-seq but none for scRNA-seq, such as the Gene Co-Expression Analysis Method. If we compare the gene expression matrix of traditional bulk RNA-seq and scRNA-seq, the gene expression matrix of traditional RNA sequencing has less noise and fewer missing values. For scRNA-seq, the experimental procedures are more sophisticated, resulting in greater noise. At the same time, the depth of sequencing is shallower, producing more missing values. Because the two RNA sequencing methods are the same, this study tried to reduce the noise and missing values of the scRNA-seq data to mimic traditional bulk RNA-seq data using the algorithm way.

We can generate ensembled data for each scRNA-seq sample, and every sample has the same attributes with bulk RNA-seq, include noise, missing value decrease and sample's correlation, low expression gene value increase.

In conclusion, our method can use in diverse situations, like comparing existing bulk RNA-seq data and scRNA-seq. From the perspective of analysis, we can generate a lot of purely bulk RNA-seq samples, such as turn cancer cells into pure cancer tissue.

Contents

List of Figures

List of Tables

1. Introduction

Nowadays, single-cell RNA sequencing is mass release. That is, if we can more use single-cell RNA sequencing datasets, we can figure out more detail about cell type identification and heterogeneity of cell responses. However, sometimes single-cell RNA sequencing data is limited because single-cell RNA sequencing data have more missing value and noise, and these two attributes cause some algorithms or methods to not work well.

In the past, there were some studies [1] that add up single-cell RNA sequencing samples to mimic bulk RNA sequencing samples called pseudo bulk [2]. And another way, to reduce single-cell RNA sequencing noise, some studies will do data impute [3]. For instance, imputation tools have been proposed to estimate dropout events and de-noise data. The performance of these imputation tools is often evaluated, or fine-tuned, using various clustering criteria based on ground-truth cell subgroup labels. We provide a more reasonable method to do pseudo bulk through statistics and analysis.

Our proposal provides a method using K-nearest neighbors (KNN) and Gaussian distribution to reduce the missing values and noise in single-cell RNA sequencing samples, while can observing low gene expression, and consider that read counts are different between single-cell RNA sequencing and bulk RNA sequencing, our ensembled data have adjusted the reads, make ensembled data more like bulk RNA sequencing. So future research can use single-cell RNA sequencing data to generate pure pseudo bulk data.

2. Methods

2.1 Overview

We propose a method based on K nearest neighbors and Gaussian distribution. Bulk RNA sequencing sequences a tissue, and single-cell RNA sequencing sequences a cell. We can imagine that when many cells cluster together, it will become a tissue, and bulk RNA sequencing and single-cell RNA sequencing is based on the same sequence technical. So we provide a method to transfer single-cell sequencing data to bulk RNA sequencing data.

In the experimental stage, we collect datasets that have bulk RNA sequencing and single-cell RNA sequencing at the same time and select common genes and preprocess with scanpy [4]. And then, we get a gene expression matrix about bulk RNA sequencing and single-cell RNA sequencing. Because we used scanpy to preprocess, we can extract the nearest neighbors list from the data structure. We observe that most biological phenomena are following Gaussian distribution, so we duplicate sampling samples according to the distance apply to Gaussian distribution. Our method can provide an ensembled sample for every single-cell sample. We also performed data visualization and classifier methods to verify that the data we simulated is credible. An overview of our proposed method is depicted in Figure 1.

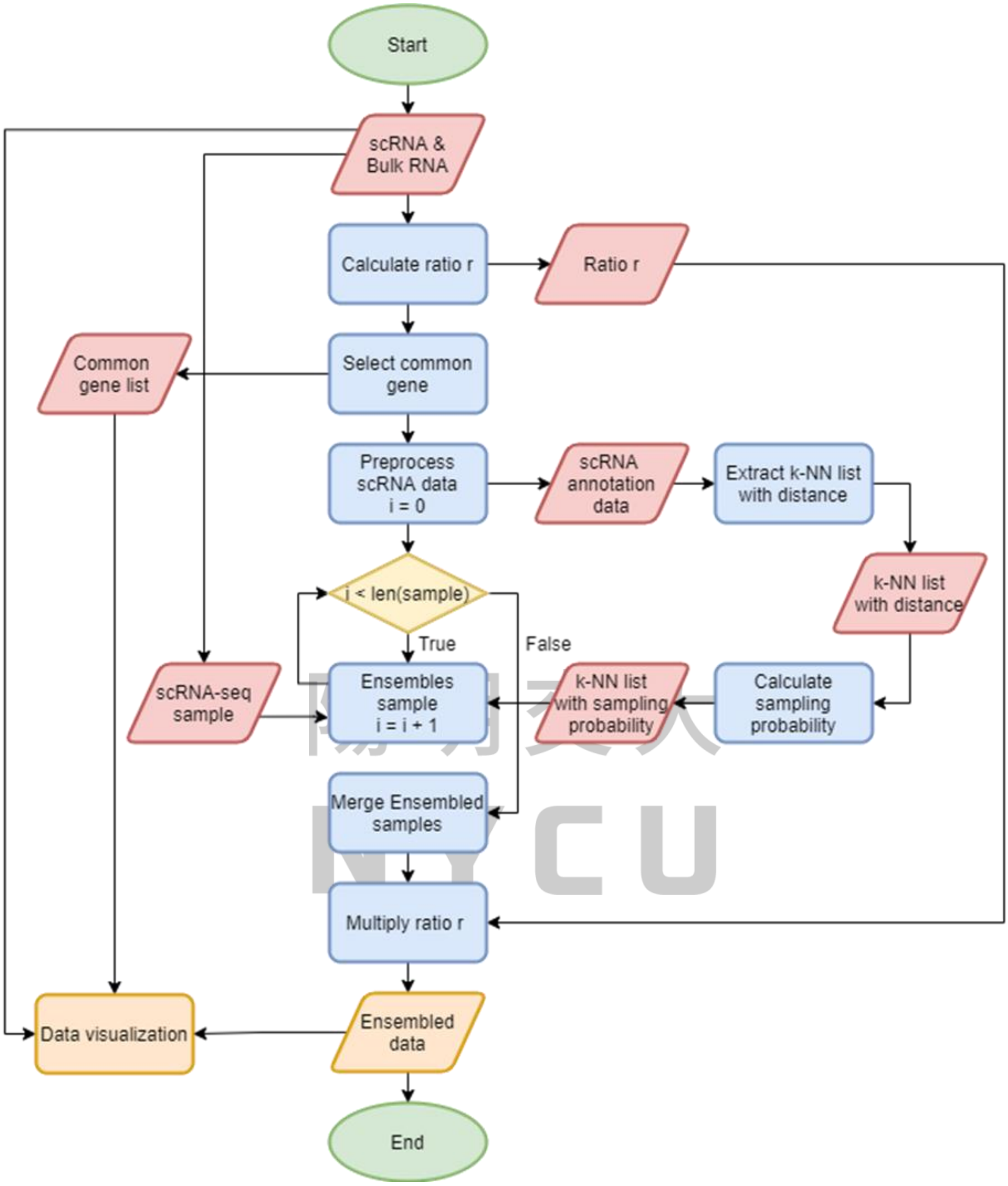

Figure 1 | Flowchart of the approach.

Input single-cell RNA sequencing and bulk RNA sequencing data, ensemble every sample of single-cell RNA sequencing, and output ensembled data.

2.2 Data

2.2.1 Attribute difference in bulk RNA-seq and scRNA-seq

scRNA-seq's attributes have pros and cons. The advantage is compared to bulk RNA-seq, scRNA-seq can analyze specific cells and more samples; for example, in dataset GSE141834 [5], they have 2400 scRNA-seq samples but only 18 bulk RNA-seq samples. The disadvantage is sparseness and more noise - sparseness means scRNA-seq's gene expression matrix has more missing values and noise means the scRNA-seq experiment is more complex.

2.2.2 Datasets

For the dataset, we need to find a data set that has both single-cell RNA sequencing and bulk RNA sequencing, so that the data visualization and parameter adjustment can be carried out better.

We chose the dataset GSE141834 and GSE136148 [6] on GEO [7]. As said before, they provide single-cell RNA sequencing data and bulk RNA sequencing data at the same time. GSE141834 contains the expression profiles of human breast cancer cells in response to glucocorticoids, a class of steroid hormones. The bulk RNA-seq data collected from populations of cells can give a detailed picture of the global transcriptional hormone response and the scRNA-seq data allow to examine the glucocorticoid response in individual cells [5]. GSE136148 provides Single-cell RNA sequencing and bulk RNA sequencing data of human breast cancer cells and mouse mammary glands [6].

According to the description of the dataset, they provide a normalization gene expression profile, so we run scanpy for preprocess. It can filter mitochondrial genes and outliers and calculate the distance of sample-sample in high dimensions. At the last, we label samples for data visualization.

2.2.3 Data visualization

We visualize the single-cell data to determine the distribution of the overall data because the KNN algorithm calculates the distance between samples in vector space.　Figuring out the distribution of the sample can evaluate the effect of the KNN graph.

2.3 K-nearest neighbors

K is a constant to evaluate target samples type or value. In KNN classification, target samples type will be decided from the most class in k nearest neighbors. For example, when the value k is 5, the three nearest neighbors' type is A, and the two nearest neighbors' type is B, the target samples type will be A. In KNN regression, the output is the value of the target sample, and this value will be an average of k nearest neighbors.

According to the KNN algorithm, closer samples will be more similar, so we just figure out the K-nearest neighbors of the target sample to use our method, such as Figure 2. We can use the below function to calculate the distance between two samples.

$$\mathrm{d}(s_i, s_j) = \sqrt{\sum_{k=1}^{n} (s_{ik} - s_{jk})^2}$$

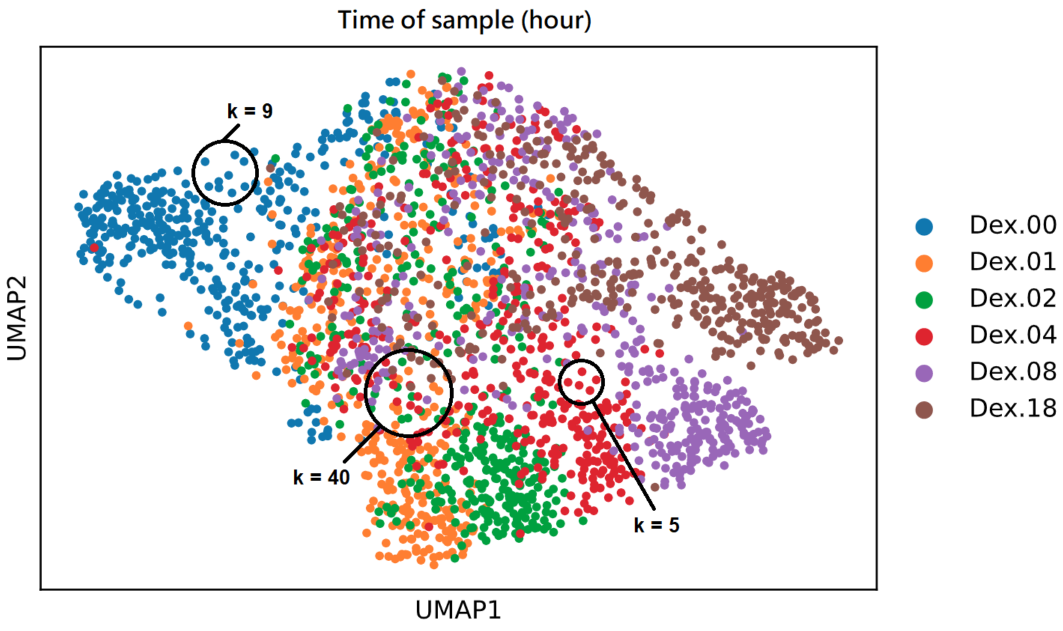

Figure 2 | K-nearest neighbors graph clear present.

If the value of K is too large, it will include too many different cell-type.

2.4 Duplicate sampling

In our method, after we have the K-nearest neighbors' list, to ensemble scRNA-seq sample to bulk RNA-seq sample, we need to duplicate sampling and adjust read counts difference between two data types. Our method provides two solutions:

1. Gaussian distribution sampling
2. Average sampling

Gaussian distribution sampling is applying sample-sample distance to Gaussian distribution because many biological phenomena follow Gaussian distribution. The probability density function is below:

$$f(x) = \frac{\exp\left(-\frac{x^2}{2}\right)}{\sqrt{2\pi}}$$

After we have the probability of every nearest neighbor, we normalize each value to use this function:

$$p' = \frac{p}{\sum_{k=1}^{n} p_k}$$

At the last, using p' to resampling gene expression and ensemble nearest neighbors sample value to output ensembled data.

2.5 Validation

2.5.1 Logistic regression

Logistic regression is a statistical model to predict variables' relationships. Logistic regression allows us to estimate how the dependent variable changes as the independent variable changes. A binary logistic model has a dependent variable with two possible values. For example, True/False will be represented by 1 and 0.

2.5.2 Linear Support Vector Machine

Linear SVM [8] is used for linearly separable data, which means a dataset can be classified into different classes by using straight lines in vector space. This data is called linearly separable data, and the classifier used is called a Linear SVM classifier.

In the linear SVM (Support Vector Machines) part, we also use the same package scikit-learn. We use the sklearn.svm.LinearSVC function,. The two parameters X and y remain the same and use the gene expression profile and sample type too.

2.5.3 Cross-validation

Machine learning requires big data to train models, but sometimes we do not have enough data, so we need to do cross-validation. Suppose a model that just repeats the labels of the samples that it has just seen. This would have a perfect score but fail to predict anything useful on unseen data. This situation is called overfitting, so the training and testing set should be separate. To avoid it, we use the function *train_test_split()*, to randomly split training and test sets. And we use function *cross_val_score()* to do cross-validation. Figure 3 is a schematic diagram to visually show this. First, the training and test data are separated. During training, we split training data into a different split and take turns with the training and test data so that we can get a more confident model.

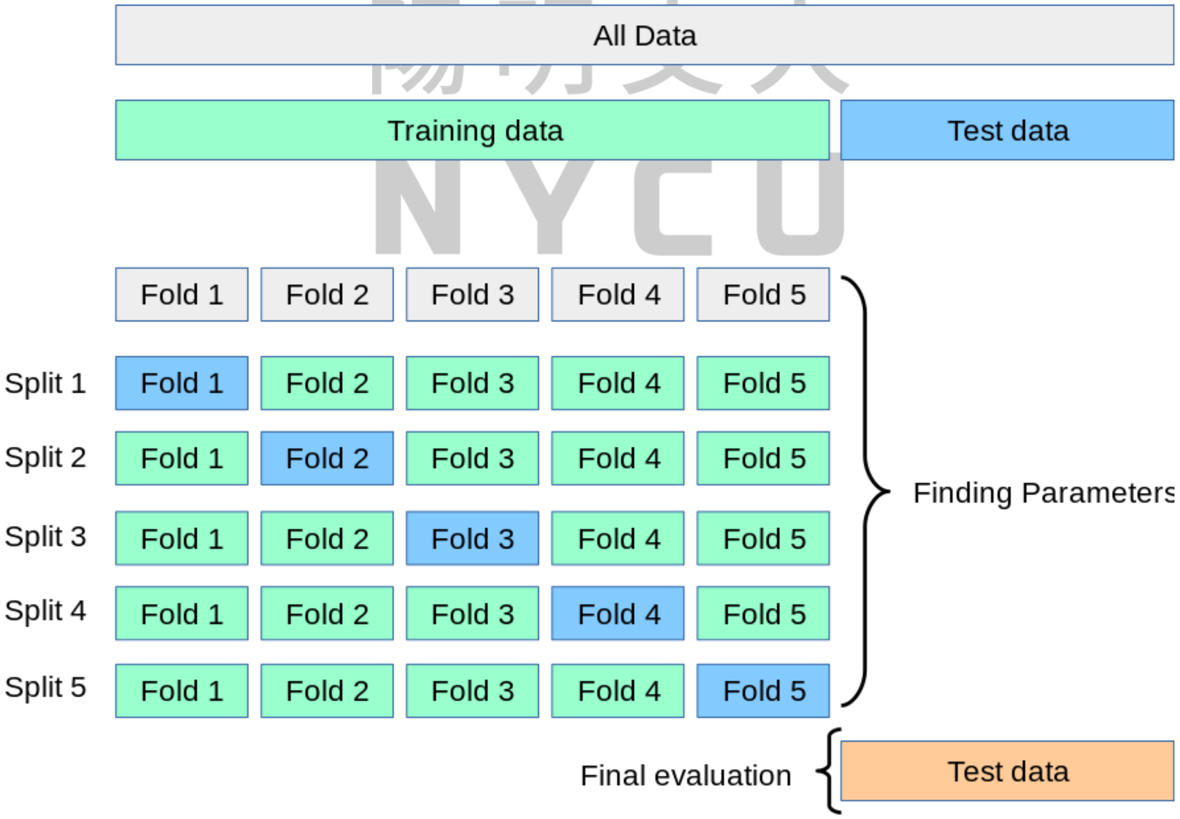

Figure 3 | Cross validation.

Split training data into a different split and take turns as the training and test data.

2.5.4 Weighted gene co-expression network analysis

Weighted gene co-expression network analysis (WGCNA) [9] is a systems biology method for describing the correlation patterns among genes across microarray samples. Weighted correlation network analysis can summarize high correlation genes together, called gene modules, those gene modules can be present by eigengene or hub gene, using eigengene or hub gene can research diverse biological topics. For example, investigate module-module interaction or overlap can analyze cancer, mouse genetics, yeast genetics, and brain imaging data. WGCNA also can be used to identify candidate biomarkers or therapeutic targets.

3. Results

3.1 Characterization of single-cell and bulk RNA-seq data

We analyze dataset GSE141834' scRNA-seq and bulk RNA-seq sample and confirm that scRNA-seq samples are more sparse. At the same time, we confirm scRNA-seq samples have a pattern in a high-dimensional space. Figure 4 reveals that bulk RNA-seq's gene expression profiles attributes are different from scRNA-seq. Figure 5 shows off scRNA-seq has bigger noise because gene' read counts standard deviation is larger. In Figure 6, we find out the sample does change with the time of the experiment.

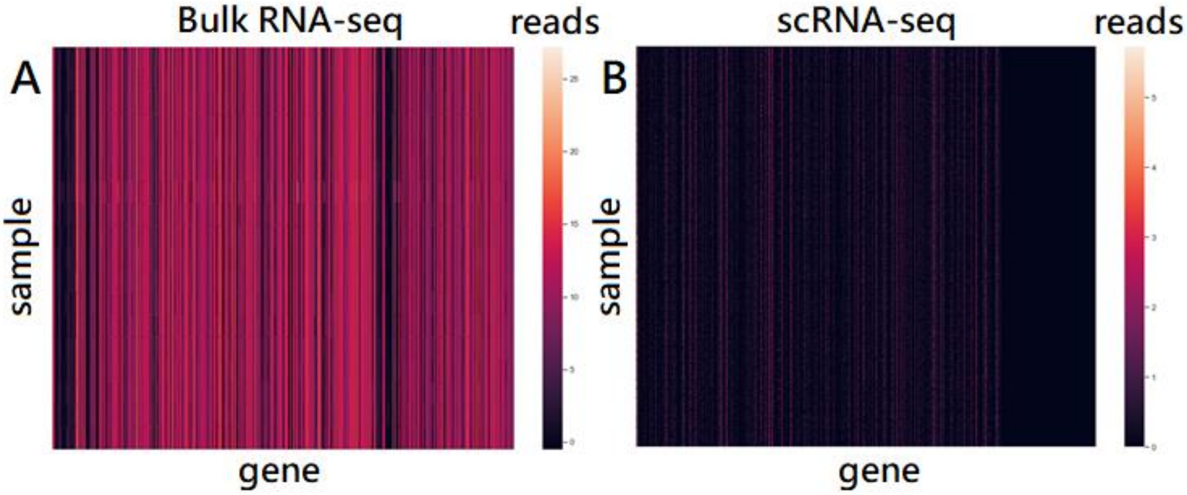

Figure 4 | Comparing gene expression profiles.

(A) the bulk RNA-seq gene expression profile, there have 18 bulk RNA-seq samples. (B) scRNA-seq gene expression profile, there have 2400 scRNA-seq samples. Both are from the dataset GSE141834.

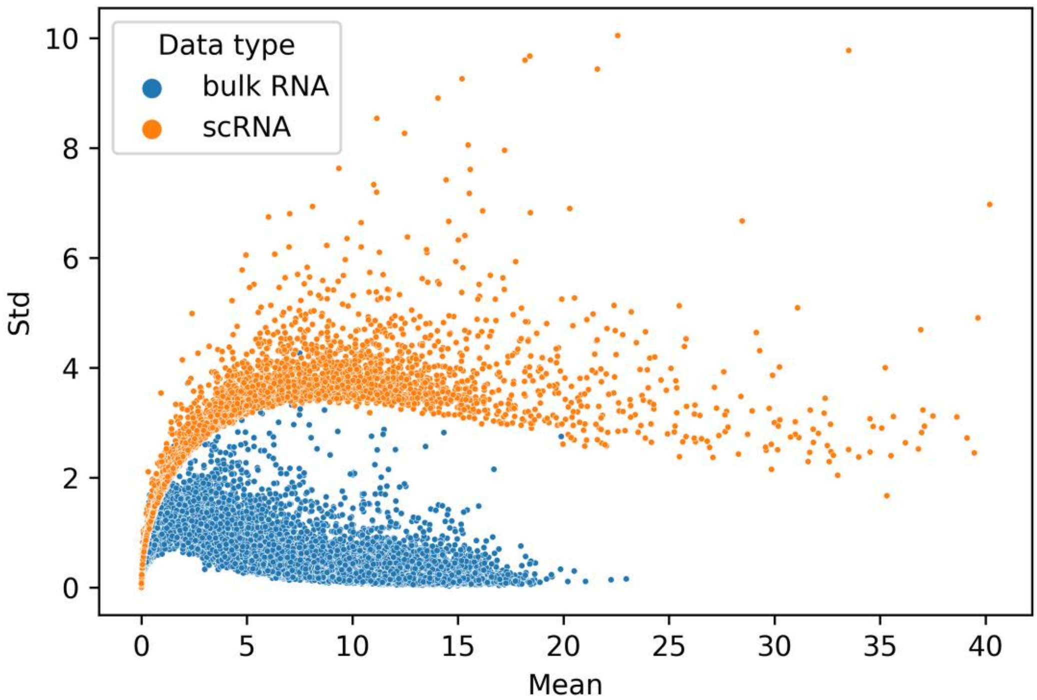

Figure 5 | View the standard deviation of gene expression.

Each point presents one gene. Bulk RNA-seq has 18 samples and scRNA-seq has 2400 samples. Bulk RNA-seq gene expression standard deviation is low. Relatively speaking, single-cell RNA-seq standard deviation is very high, which means that single-cell RNA-seq' noise is bigger than bulk RNA-seq.

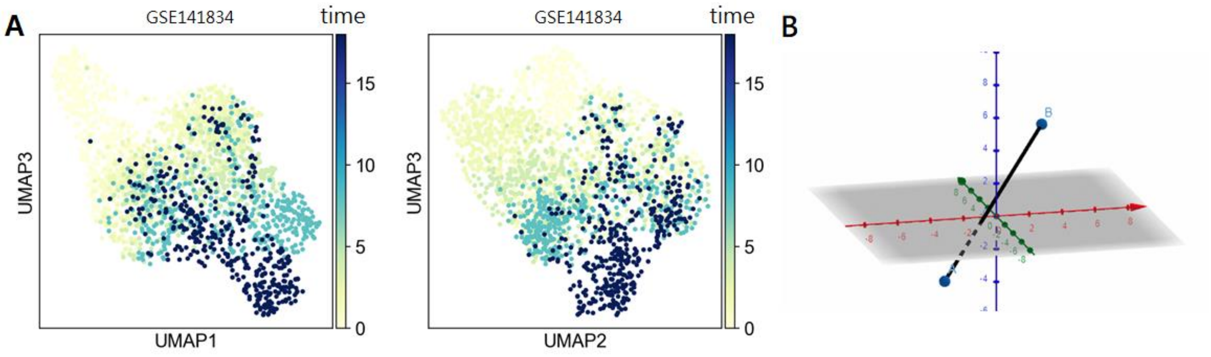

Figure 6 | Data visualization.

(A) Projections of the first three dimensions for the UMAP representation of the scRNA-seq samples in dataset GSE141834. Each point represents one sample (cell). (B) The samples have a linearly gradient pattern in three-dimensional space.

3.2 Ensembled data reduce missing values

Ensembled data can reduce missing value compared with scRNA-seq data because our method integrates k nearest neighbors gene expression profile. Figure 7 and Figure 8 reveals the missing value in ensembled data is a significant decline. Take dataset GSE141834 as an example, each sample reduced 5000 missing values on average. The fewer missing values, the better to run WGCNA.

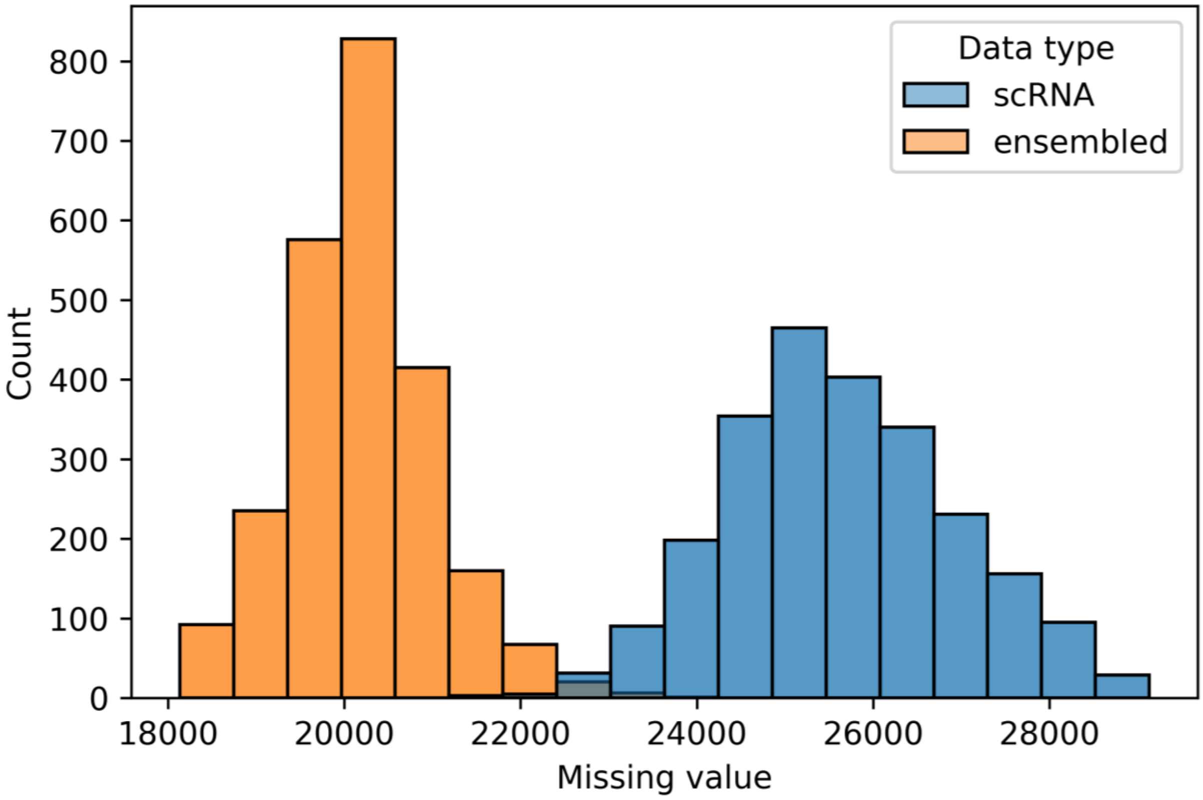

Figure 7 | Histogram of the dataset GSE141834' ensembled data and scRNA-seq data' missing value comparing.

Ensembled data' missing values (per sample) are between 18,000 to 24,000 and scRNA-seq data' missing values are between 21,000 to 29,000. The sample size of both is 2400.

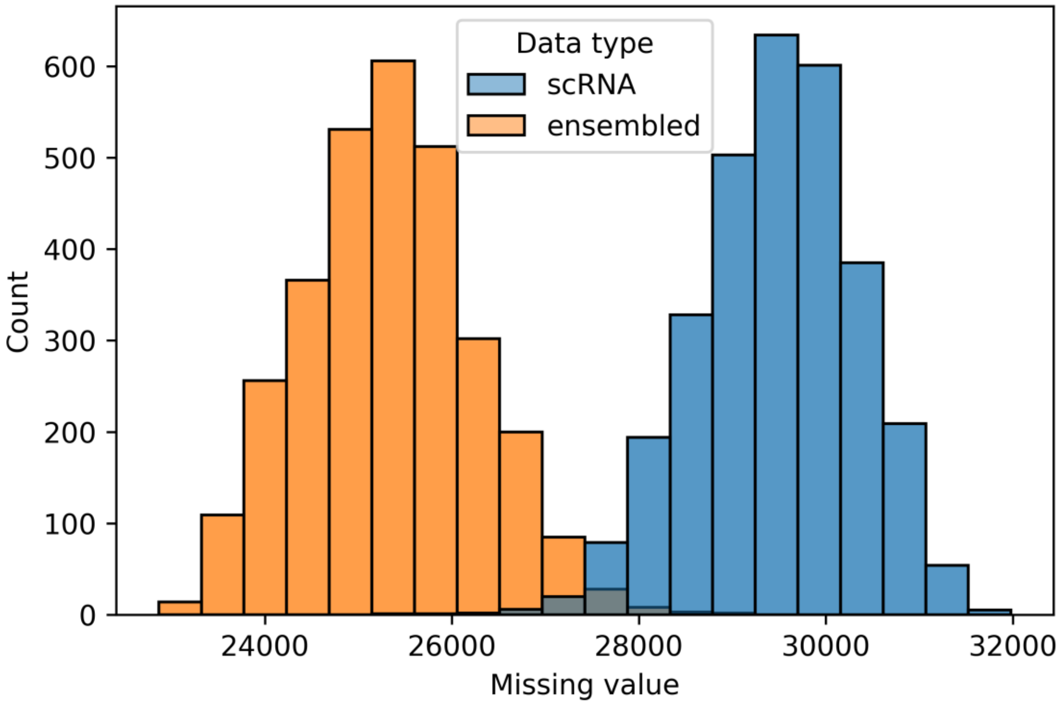

Figure 8 | Histogram of the dataset GSE136148' ensembled data and scRNA-seq data' missing value comparing.

Ensembled data' missing values (per sample) are between 22,000 to 29,000 and scRNA-seq data' missing values are between 26,000 to 32,000. The sample size of both is 3022.

3.3 Ensembled data reveal lowly expressed genes

Ensembled data can observe low gene expression. Through our method and adjusted gene expression profile by ratio r, we can get samples extremely similar to Bluk RNA-seq. This is better looking for weak gene expression in scRNA-seq samples. Figure 9 and Figure 10 can observe that gene expression is increased. For instance, in dataset GSE141834 the average number of genes in each sample has increased from 6,500 to 11,000, which is an increase of 70%.

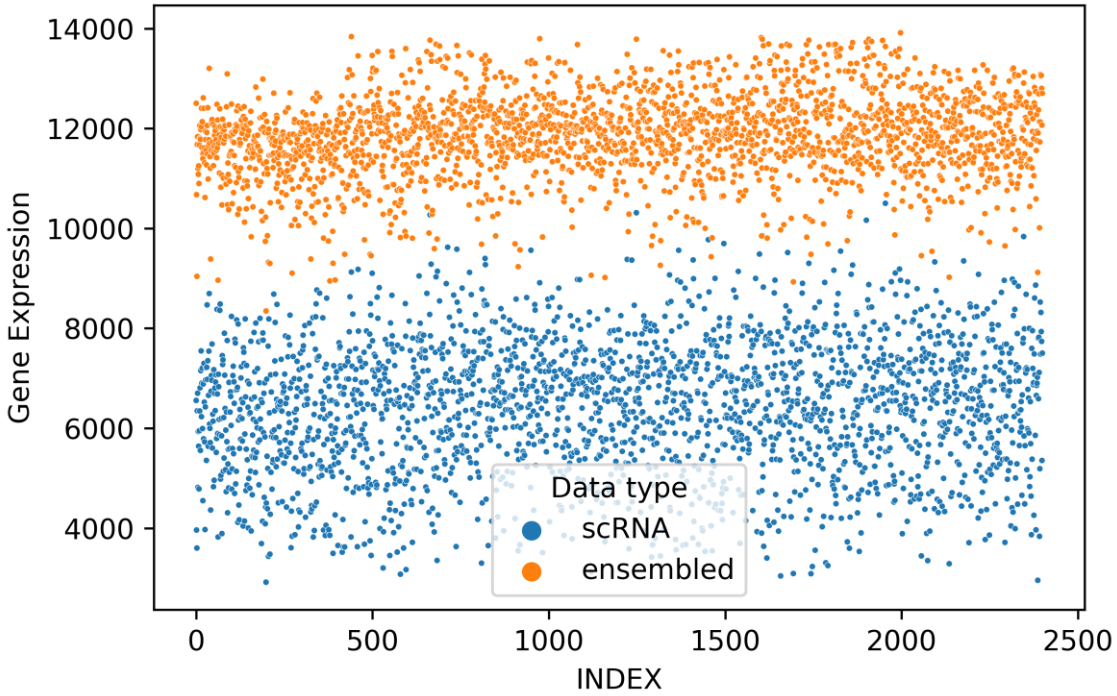

Figure 9 | Scatter plot of the dataset GSE141834' ensembled data and scRNA-seq data' gene expression.

Ensembled data' gene expression (per sample) are between 8,000 to 14,000 and scRNA-seq data' gene expression are between 2000 to 11,000. The sample size of both is 2400.

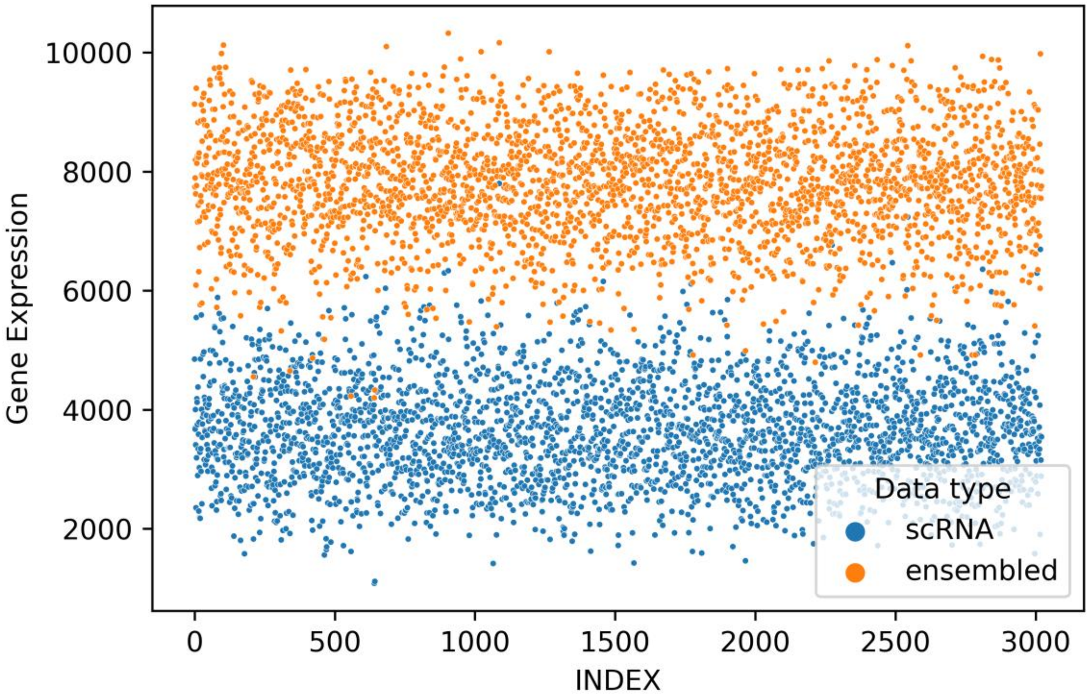

Figure 10 | Scatter plot of the dataset GSE136148' ensembled data and scRNA-seq data' gene expression.

Ensembled data' gene expression (per sample) are between 8,000 to 14,000 and scRNA-seq data' gene expression are between 2000 to 11,000. The sample size of both is 3022.

3.4 Ensembled data can be seen as bulk RNA-seq data

As you can see from Figure 11, the average number of reads for a single-cell sample is less than 10000, but the average number of reads per sample for the bulk-RNA-seq data and the data we synthesized is the same range of 170000 and 250000. The same situation can observe from Figure 12, ensembled data and bulk RNA-seq data' read counts range are the same.

Coupled with the previous verification of properties, the data we ensemble can be seen as bulk RNA-seq data, which can be adapted as a method designed for bulk RNA-seq.

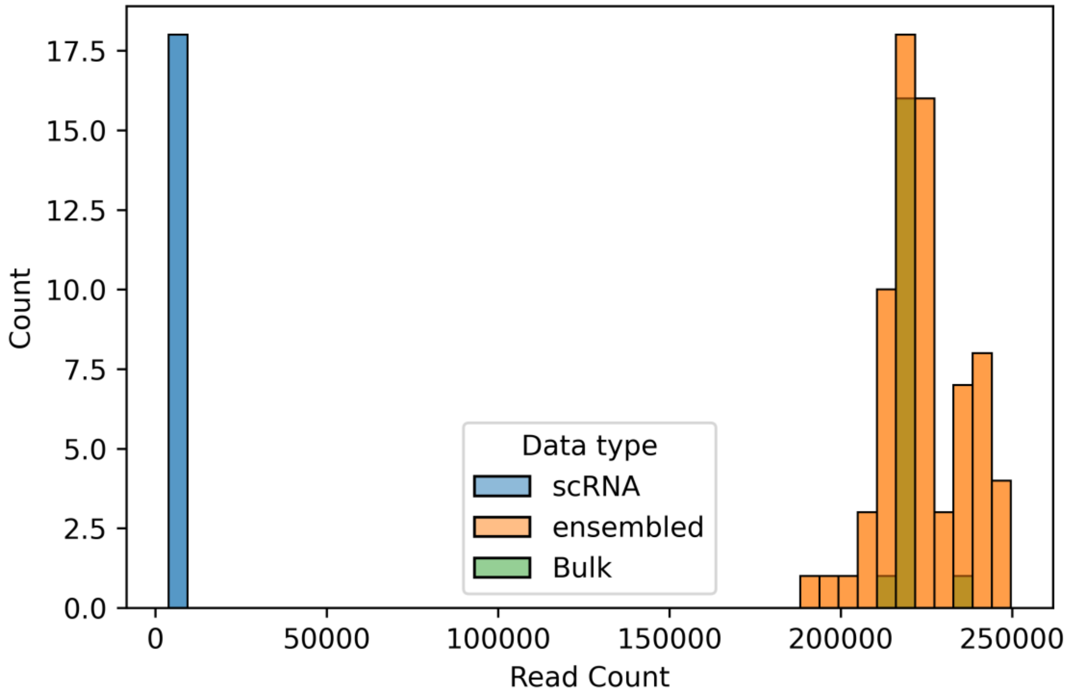

Figure 11 | Histogram of the dataset GSE141834' ensembled, scRNA-seq, and bulk RNA-seq data' read counts.

Ensembled and bulk RNA-seq data's read counts (per sample) are between 170,000 to 250,000 and scRNA-seq data's read counts are less than 10,000.

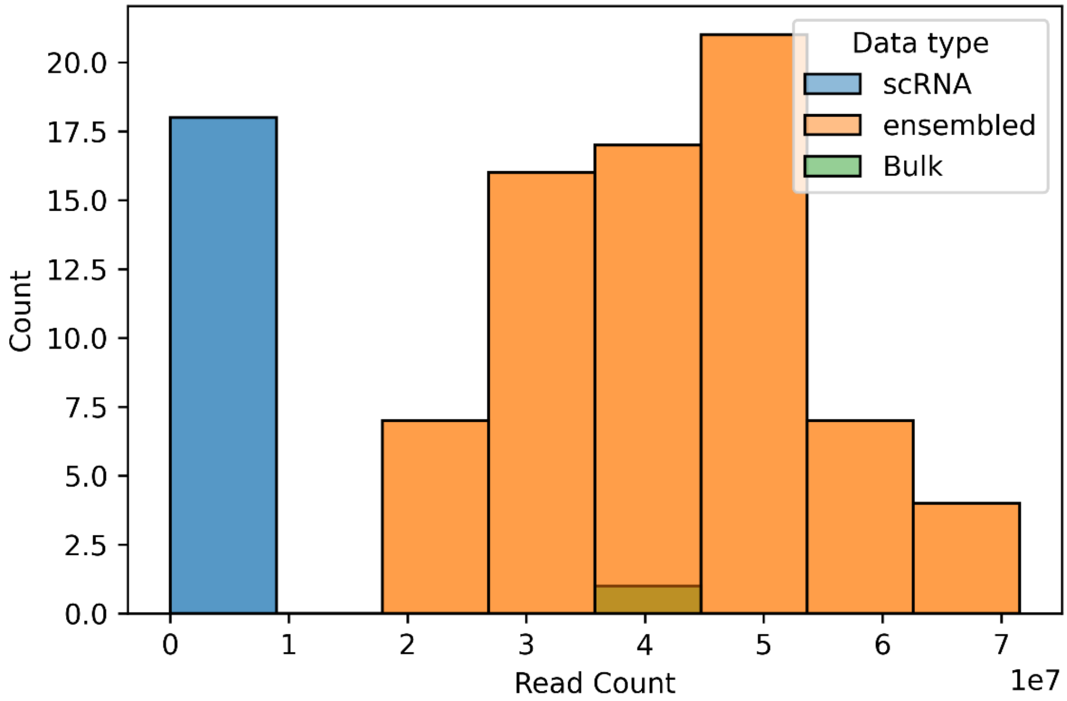

Figure 12 | Histogram of the dataset GSE136148' ensembled, scRNA-seq, and bulk RNA-seq data' read counts.

Ensembled and bulk RNA-seq data's read counts (per sample) are between 2 * 10^7 to 8 * 10^7 and scRNA-seq data's read counts are less than 1 * 10^7.

3.5 Ensembled data decrease scRNA-seq noise

We described that scRNA-seq noise is bigger than bulk RNA-seq before. The larger the standard deviation, the more noise. Our ensembled data can output denoise data. In Figure 13 and Figure 14, we can observe ensembled data that keep the scRNA-seq attributes and reduce the standard deviation.

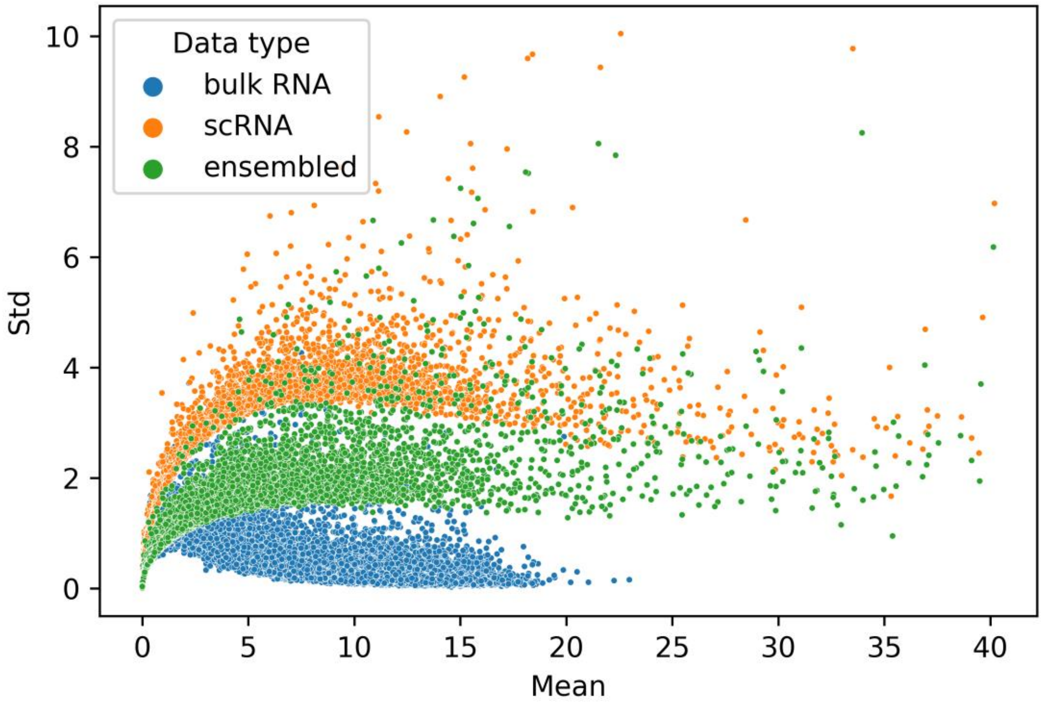

Figure 13 | Scatter plot of the dataset GSE141834' ensembled, scRNA-seq, and bulk RNA-seq gene expression' mean and standard deviation.

To compare with ensembled and bulk RNA-seq data, scRNA gene expression values have been adjusted by multiply ten times. scRNA-seq and ensembled data have 2400 samples, and bulk RNA-seq has 18 samples.

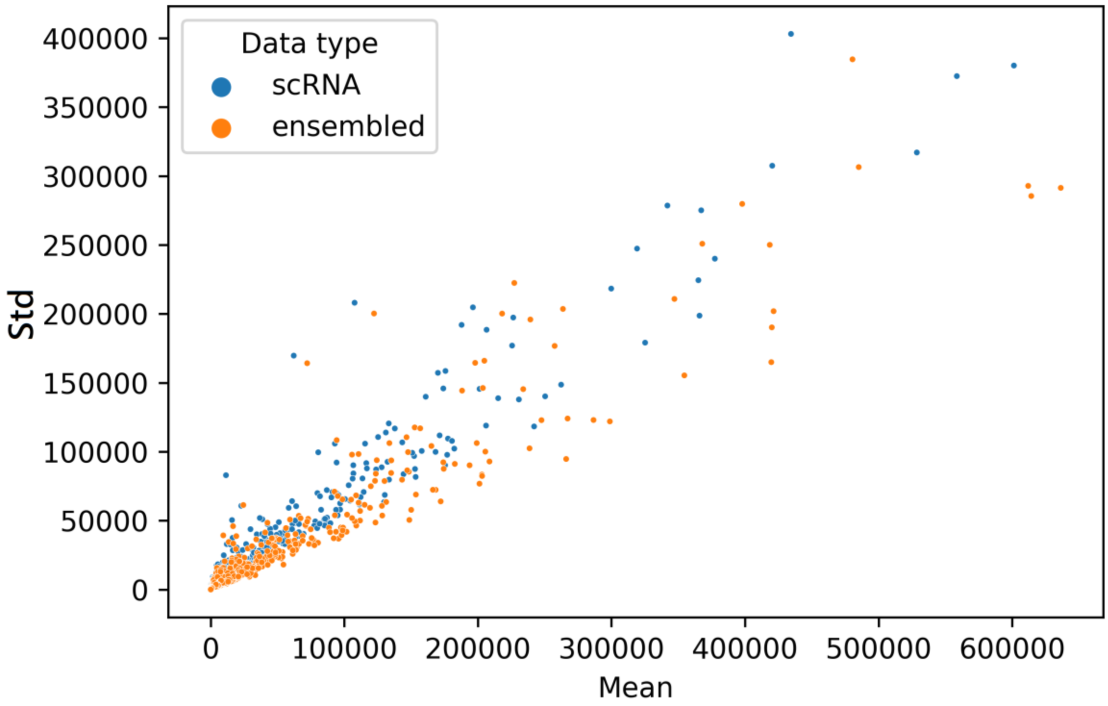

Figure 14 | Scatter plot of the dataset GSE136148' ensembled and scRNA-seq gene expression' mean and standard deviation.

To compare with ensembled and bulk RNA-seq data, scRNA gene expression values have been adjusted by multiply ten times. The p-value between the two is 3.3168935076198046 * 10^{-58}.

3.6 Validation of different sampling methods

In our validation, we use *sklearn.linear_model.LogisticRegression().fit(X, y)* function in python package called scikit-learn. Variable *X* is gene expression profile and *y* is sample type. Through Logistic regression, we validate that the dataset GSE141834 Gaussian distribution sampling is better than average sampling because Gaussian distribution sampling ensembled data have 83% classification accuracy, and average sampling ensembled data have 79.5% classification accuracy.

Table 1 | Logistic regression classification accuracy of two different sampling.

Average accuracy of each cv split	Gaussian distribution	Average sampling
Logistic Regression	83%	79.5%

Linear SVM reveals that in dataset GSE141834, Gaussian distribution sampling is better than average sampling too, with 76% classification accuracy of Gaussian distribution sampling and 74.5% classification accuracy of average sampling.

According to logistic regression and linear SVM results, we realize that ensembled data are preserved scRNA-seq sample's class attribute.

Table 2 | Linear SVM classification accuracy of two different sampling.

Average accuracy of each cv split	Gaussian distribution	Average sampling
Linear SVM	76%	74.5%

3.7 Gene co-expression network analysis with ensembled data

We performed gene co-expression network analysis with ensembled data using WGCNA [9] and compared the results with those from the original scRNA-seq data. Running WGCNA requires calculating the correlation between samples. The scRNA-seq samples contain many missing values that make the correlation calculation very difficult. As shown in Figure 15, our ensembled data preserve the scRNA-seq sample attributes. In the 0-hr and 1-hr conditions, two modules with their hub genes as TOP2A and RPL8, respectively, were identified from both data. In the 2-hr condition, since our ensemble method reduces the missing values to mimic bulk RNA-seq data, we can successfully find out the gene co-expression modules from the ensembled data, while the original scRNA-seq data are failed.

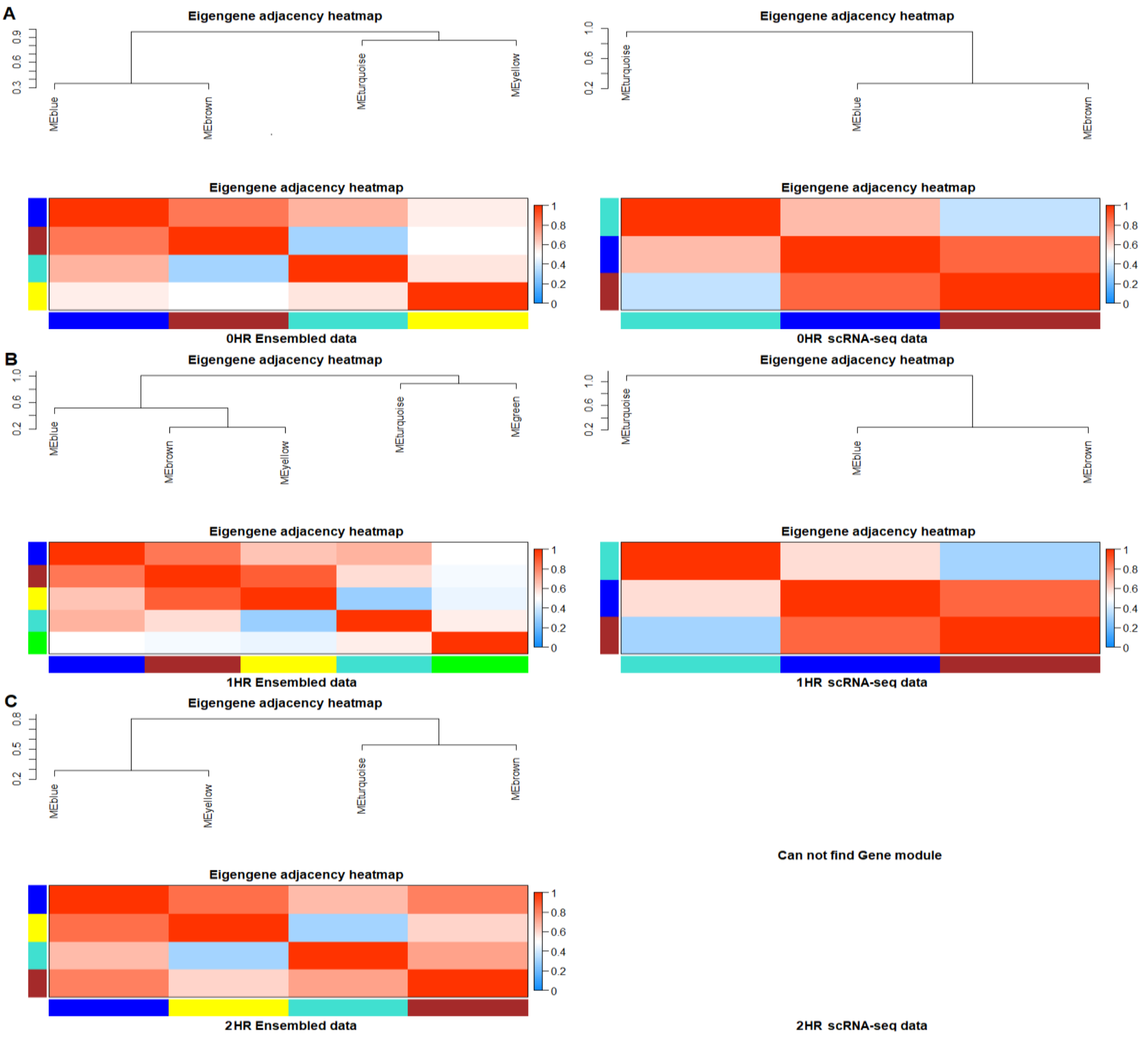

Figure 15 | Ensembled and scRNA-seq data run on WGCNA.

(A, B) Our ensembled data can find more gene modules in a default setting, overlap hub genes are TOP2A and RPL8. (C) original scRNA-seq data cannot find any gene module because the missing value is too much.

4. Discussion

We mention that some research will add up scRNA-seq samples to mimic bulk RNA-seq samples. Using our method can provide more reliable and quantity bulk-liked samples because we can output ensembled data for each scRNA-seq sample.

A limitation is that currently scRNA-seq usually have insufficient sequencing depth, so our ensembled data cannot perfectly mimic bulk RNA-seq data. If future scRNA-seq sequencing depth can go deeper, our method can increase higher accuracy.

Future research can focus on how to find out suitable k-value for each dataset. For example, we can use data set volume, sample distance, and cell type to figure out the pattern of the k-value.

5. Conclusion

Our method is especially suitable for cancer tissue because the sample of bulk RNA-seq on cancer usually mixes different cell-type. We can integrate cancer cells and mimic them to purely bulk-liked samples.

In conclusion, our method output ensembled data, and that have attributes just like bulk RNA-seq data. For example, ensembled data have few missing values, denoised, and more connective than scRNA-seq data, compare with bulk RNA-seq, ensembled data are more purely. scRNA-seq data can use our method to mimic bulk RNA-seq to find more detail in gene expression profiles.

6. References

1. Mengqi Zhang, Si Liu, Zhen Miao, Fang Han, Raphael Gottardo, Wei Sun. 2021. Individual Level Differential Expression Analysis for Single Cell RNA-seq data. bioRxiv doi: https://doi.org/10.1101/2021.05.10.443350
2. Hao Y, Hao S, Andersen-Nissen E, III WMM, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zagar M, Hoffman P, Stoeckius M, Papalexi E, Mimitou EP, Jain J, Srivastava A, Stuart T, Fleming LB, Yeung B, Rogers AJ, McElrath JM, Blish CA, Gottardo R, Smibert P, Satija R (2021). "Integrated analysis of multimodal single-cell data." Cell. doi: 10.1016/j.cell.2021.04.048, https://doi.org/10.1016/j.cell.2021.04.048.
3. Feng X, Chen L, Wang Z, Li SC. I-Impute: a self-consistent method to impute single cell RNA sequencing data. BMC Genomics. 2020 Nov 18;21(Suppl 10):618. doi: 10.1186/s12864-020-07007-w. PMID: 33208097; PMCID: PMC7677776.
4. Wolf, F., Angerer, P. & Theis, F. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol 19, 15 (2018). https://doi.org/10.1186/s13059-017-1382-0
5. Hoffman, J.A., Papas, B.N., Trotter, K.W. et al. Single-cell RNA sequencing reveals a heterogeneous response to Glucocorticoids in breast cancer cells. Commun Biol 3, 126 (2020). https://doi.org/10.1038/s42003-020-0837-0
6. Meichen Dong, Aatish Thennavan, Eugene Urrutia, Yun Li, Charles M Perou, Fei Zou, Yuchao Jiang, SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references, Briefings in Bioinformatics, Volume 22, Issue 1, January 2021, Pages 416–427, https://doi.org/10.1093/bib/bbz166
7. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002 Jan 1;30(1):207-10. doi: 10.1093/nar/30.1.207. PMID: 11752295; PMCID: PMC99122.
8. Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., …

others. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830.

9. Langfelder, P., Horvath, S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559 (2008). https://doi.org/10.1186/1471-2105-9-559