News Release

Title

Development of Artificial Intelligence Technology to Extract Useful Knowledge from Single-Cell Multi-Omics Data
~Accelerating the understanding of diseases at the single-cell level~

Key Points

- The group have developed an artificial intelligence technology, scMM, which compresses and integrates the information of multiple modalities in single-cell multi-omics data using deep generative models, a kind of deep learning.
- The scMM automatically learns the relationships among modalities in single-cell multiomics data and successfully predict the missing modalities by transforming the information across modalities.
- The scMM can effectively extract useful knowledge from single-cell multiomics data, and is expected to accelerate research using single-cell multiomics analysis.

Summary

A research group led by Kodai Minoura, a medical student, Professor Teppei Shimamura of the Division Systems Biology, Nagoya University Graduate School of Medicine (Dean: Kenji Kadomatsu), and Professor Hiroyoshi Nishikawa of the Division of Immunology, Nagoya University Graduate School of Medicine, has successfully developed an artificial intelligence technology that can extract useful knowledge from single-cell multiomics data by applying deep generative model. Single-cell analysis technology, which has made remarkable progress in recent years, is now capable of comprehensively measuring information on modalities such as the transcriptome, epigenome, and cell surface markers at the single-cell level, and is actively used to analyze the diversity of disease-causing cells and the functions of individual abnormal cells. In particular, single-cell multiomics analysis, in which multiple modalities are measured simultaneously from the same cell, is now attracting attention. Single-cell multiomics analysis is expected to reveal the diversity and functions of cells that cannot be captured by a single modality. However, single-cell multiomics data is big data containing complex information from multiple modalities, and methods for discovering useful medical-biological knowledge from such data have been limited. To address this issue, the research group developed scMM (a mixture-of-experts deep generative model for integrated analysis of single-cell multi-omics data), an artificial intelligence technology specialized for the analysis of single-cell multiomics data. scMM is based on a deep generative model capable of inferring the latent state of individual data points from large-scale data, and enables fully automated integration and compression of multiple modalities and discovery of relationships between modalities. This research is expected to accelerate research using single-cell

multiomics analysis and contribute to a better understanding of diseases and the establishment of new therapies. The results of this study were published online in Cell Reports Methods (September 15, 2021).

Research Background

Recently, a series of technologies have been developed to measure information such as the transcriptome (single-cell RNA-seq; scRNA-seq), cell surface markers, and chromatin accessibility (single-cell Assay for Transposase Accessible Chromatin; scATAC-seq) at the single-cell level, contributing to important discoveries in various fields such as cell diversity in normal tissues, cancer heterogeneity, and cell fate tracking. In addition, while the single-cell analysis technology used to measure a single modality, now with the advancement of technology, it is possible to simultaneously measure multiple modalities such as transcriptome and epigenome, transcriptome and cell surface markers for each cell. The data obtained by such single-cell multiomics analysis is expected to reveal the diversity of cells and relationships among modalities that cannot be captured by conventional single-modality analysis. On the other hand, single-cell multiomics data contains very high-dimensional data for each cell and has a complex structure because the data is acquired across multiple modalities. Therefore, bioinformatics analysis methods to computationally extract useful knowledge from single-cell multiomics data are currently very limited. In this study, we developed a model with the main purpose of accurately compressing and integrating multiple modalities and supplementing missing modalities by learning the relationships among modalities.

Research Results

The scMM developed in this study is based on a deep generative model and is specialized for the analysis of single-cell multi-omics data (Fig. 1). The deep generative model has been widely applied as a very powerful model to automatically infer the parameters of the underlying probability distribution from a large amount of data. The scMM model is constructed by selecting a probability distribution that reflects the data characteristics of each modality, and successfully integrates multiple modalities in single-cell multiomics data using a method called mixture-of-experts. As a result of verifying the performance of scMM using publicly available single-cell multiomics data of scRNA-seq of human peripheral blood mononuclear cells and cell surface markers, it was found that scMM can separate cell populations with higher granularity than conventional methods by compressing the information of multiple modalities (Fig. 2). In addition, the generation of "peudocells," which takes advantage of the ability of deep generative models to generate data, makes it possible to visualize what information is contained in each dimension of the compressed information. In addition, by generating data across modalities, it was shown that it is possible to predict cell surface markers with high accuracy from scRNA-seq data alone. Similar results were obtained in validation using publicly available single-cell multiomics data of scRNA-seq and scATAC-seq.

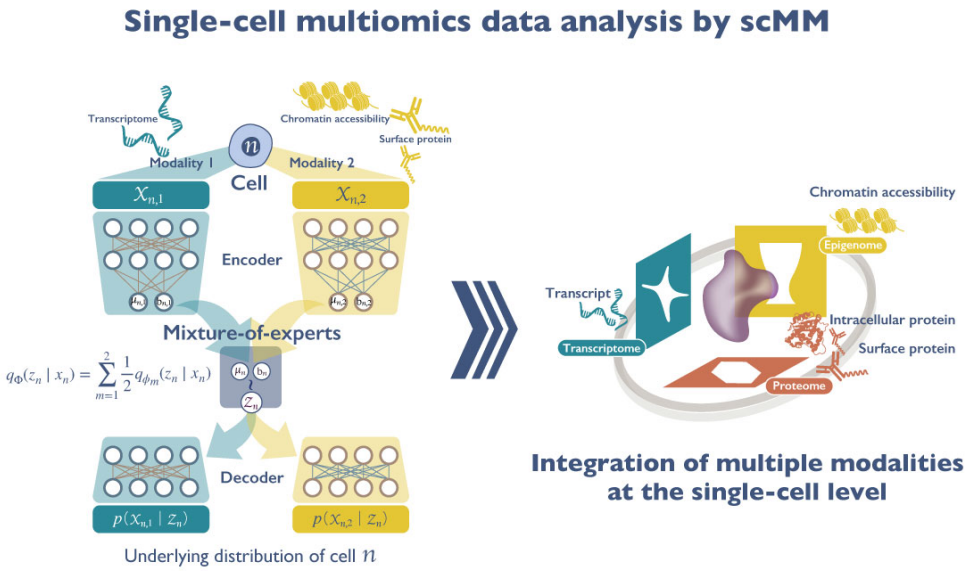

Fig. 1: A deep generative model (scMM) specific to single-cell multi-omics data enables the integration of multiple modalities at the single-cell level.

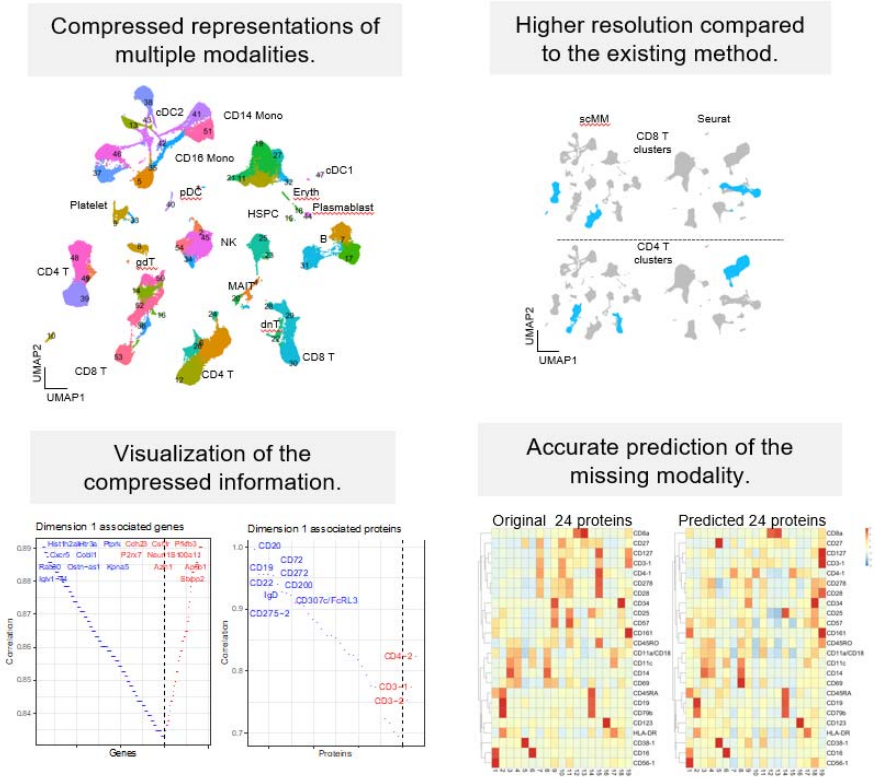

Fig. 2: Figure 2: (Upper left) Cell populations can be separated from a compressed representation of transcriptome and cell surface marker information. (Upper right) Cell population analysis with higher granularity compared to existing methods. (Bottom left) Visualization of the transcriptome and cell surface markers in the compressed representation. (Bottom right) Using a trained model, cell surface markers can be predicted with high accuracy from transcriptome information alone.

Research Summary and Future Perspective

By using a probability distribution according to the data distribution of each modality, scMM is a flexible model that can handle various modalities such as DNA methylation, histone modification, and spatial transcriptome in addition to the modalities verified in this study. It is expected that scMM will be further extended and developed as a basis for future single-cell multi-omics analysis.

Publication

Journal: Cell Reports Methods
Title: A mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data.
Author/affiliation:

Kodai Minoura[1,2], Ko Abe[3], Nam Hyunha[1], Hiroyoshi Nishikawa[2,4], and Teppei Shimamura[1,*]
[1]Division of Systems Biology, Nagoya University Graduate School of Medicine, Nagoya, Japan.
[2]Department of Immunology, Nagoya University Graduate School of Medicine, Nagoya, Japan.
[3]Laboratory of Medical Statistics, Kobe Pharmaceutical University
[4]Division of Cancer Immunology, Research Institute/EPOC, National Cancer Center, Tokyo/Chiba, Japan.

DOI : 10.1016/j.crmeth.2021.100071

Japanese ver.
https://www.med.nagoya-u.ac.jp/medical_J/research/pdf/Cell_Rep_Met_210916.pdf