A graph-based practice of evaluating collective idendities of cell clusters

Yuji Okano[1,2], Yoshitaka Kase[2], and Hideyuki Okano[2]

[1] Department of Extended Intelligence for Medicine, The Ishii-Ishibashi Laboratory, Keio University School of Medicine
[2] Division of CNS Regeneration and Drug Discovery, International Center for Brain Science, Fujita Health University

February 26, 2024

Abstract

Random sentences

Introduction

It has been more than 10 years since the birth of single-cell RNA-sequencing (scRNA-seq), and the technology now is recognized as a prominent game changer of modern molecular biology. Likewise the pioneering technology, bulk RNA-seq, scRNA-seq can observe multidimensional gene expression profiles, while it also can provide such information in single-cell-level.

⋮

In our previous research, we proposed a gene regulatory network (GRN)-based representation of cell clusters while edges of GRNs explain statistical dependencies between two genes, and demonstrated that similarity of two clusters can be defined as a quasi-pseudo-metric function d^*[1]. When the

In addition to the theoretical proposal, we applied our logic to the annotation of scRNA-seq data, and showed that the GRN-based anntation can visualize the similarities and the difference of cell clusters, which the conventional differentially expressed gene (DEG)-based manual annotation could not address due to the objectives of differential expression analysis (DEA) to summerize the biological semantics of the clusters. Althogh we introduced a theory to form GRNs based on dependencies of gene expressions, we compromised to implement the algorithm relying on the statistical test of correlation to deal with the continuity of scRNA-seq data. Considering the fact that our method's primary application is the annotation of scRNA-seq data, an effective binarization method is needed to reduce computational costs and streamline the overall time required to initiate main analyses. Furthermore, we intended to make our framework dependent on researchers' expertises on the sample domains so that the metrics of cellular identities are tailor-made for the research scopes providing necessary and sufficient resolutions. Constrary, this design made our algorithm unfriendly to users. As the legitimacy our theory needs to be validated in various cases, A semi-automated system to help users select key marker genes is desired.

⋮

Leveraging the backbone theory of GRN-based comparisons of cluster-wise cellular identities (i.e., cell classes), we implemented

⋮

To simplify the contents of this study to highlight our foci, we would not discuss any practices of designing data spaces or clustering in depth.

Results

Challenges of the framework of GRN-based methods

In this reserch, we revisited the workflow of the GRN-based annotation

Automated marker-gene suggestion

Although we intended to require experimenters to curate marker genes to use in GRNs, overly recurrsive trials to find

Dropout-based binarization

Benchmarking

Discussion

I have no idea.

Methods

GRNet Impletemtations

GO term-assisted gene selection referring Jaccard Index

$$J(A,B) := \frac{A \cap B}{A \cup B} \tag{1}$$

Jaccard Index of two sets A, B is defined as Eq. (1). We expanded this definition to pairwise comparisons of multiple elements by forming a matrix where each element is the corresponding Jaccard Index, and we named the matrix Jaccard index matrix (JIM). For example, the element in i-th row and j-th column (where $i, j, k \in \mathbb{N}$ and $i \leq k, j \leq k$), $JIM_{i,j}$, can be defined as follows when a JIM of sets $X_1, X_2, \cdots, X_k$ are considered:

$$JIM_{i,j} := J(X_i, X_j) \tag{2}$$

Especially for seed markers, sets of subscribed GO terms (let $G_1, \cdots, G_k$) and their JIM are calulated in order to set $min_{i,j}(J(G_i, G_j))$ as a threshold of biological correspondence.

⋮

For detailed method of implementation, we calculated the JIM of the related GO terms of given seed markers. We used mygene.py[2] to query the GO database, and Numpy[3] to calculate JIM.

GRNs and the evaluation function

Following our previous report[1], we computed GRNs by calculating correlations of continuous gene expression values (e.g., $log_2(RPM + 1)$) using Pgmpy[4]. In this study, we introduced

scRNA-seq data analysis

Dataset List

The scRNA-seq data we used in this research were publicly available as online resources as follows:
M1C10X: https://portal.brain-map.org/atlases-and-data/rnaseq/human-m1-10x
hFB: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE165388
PBMC3k: https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/pbmc3k
aHSPC: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE137864
BCA: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE149938

Preprocessing, dimensionality reduction, and visualization

We performed data preprocessing, dimensionality reduction, data visualization of the scRNA-seq datasets using Python packages (including Scanpy[5], Polars, Pandas[6], Numpy, Matplotlib[7], Seaborn[8]) and Julia packages.

Clustering and DE analysis

We performed leiden clustering, DE analysis using Scanpy.

Resource availability

Data availability

Not applicable

Code availability

GRNet and the analysis codes are available on GitHub (https://github.com/yo-aka-gene/grnet). Online documentation for GRNet is also provived (https://grnet.readthedocs.io).

Acknowledgements

We thank hogehoge for thorough support.

Abbreviations

DEA differential expression analysis

DEG differentially expressed gene

GRN gene regulatory network

JIM Jaccard index matrix

scRNA-seq single-cell RNA-sequencing

References

[1] Y. Okano, Y. Kase, and H. Okano, "A set-theoretic definition of cell types with an algebraic structure on gene regulatory networks and application in annotation of rna-seq data," *Stem cell reports*, vol. 18, no. 1, pp. 113–130, 2023.

[2] C. Wu, A. Mark, and A. I. Su, "Mygene. info: gene annotation query as a service," *bioRxiv*, p. 009332, 2014.

[3] C. R. Harris, K. J. Millman, S. J. Van Der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, *et al.*, "Array programming with numpy," *Nature*, vol. 585, no. 7825, pp. 357–362, 2020.

[4] A. Ankan and A. Panda, "pgmpy: Probabilistic graphical models using python," in *Proceedings of the 14th Python in Science Conference (SCIPY 2015)*, Citeseer, 2015.

[5] F. A. Wolf, P. Angerer, and F. J. Theis, "Scanpy: large-scale single-cell gene expression data analysis," *Genome biology*, vol. 19, pp. 1–5, 2018.

[6] W. McKinney *et al.*, "Data structures for statistical computing in python," in *Proceedings of the 9th Python in Science Conference*, vol. 445, pp. 51–56, Austin, TX, 2010.

[7] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[8] M. L. Waskom, "seaborn: statistical data visualization," *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021.

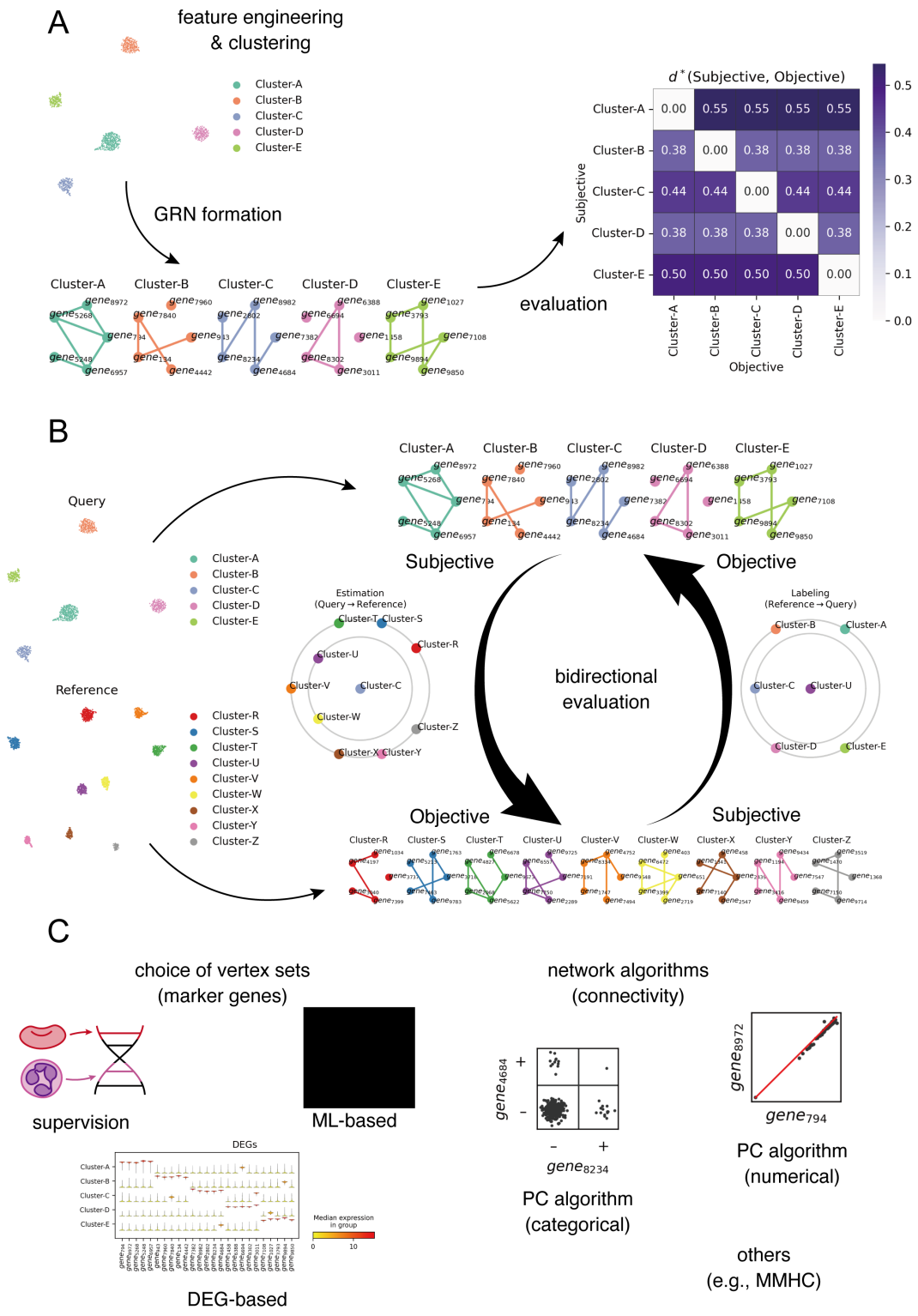

Figure 1: The framework of the GRN-based characterization and annotation of cell classes

A: The foundation of the GRN-based characterization of cell classes. After clustering in designed data space by arbitrary methods, cell classes (the clusters) can be represented by GRNs of corresponding genes of choice. The similarity of two GRNs of the same vertex (marker gene) are evaluated with the assymetrical function d^*, where the return values reflects the similarity from the viewpoint of the subjective clusters. **B**: Schematic of the GRN-based scRNA-seq data annotation. Expecting the referential data to reflect canonical states of target sample domains, the evaluation of the similarity among cell classes can be performed bidirectionally. **C**: Methodological variations of the selection of vertex sets (marker genes) and the algorithms to compute the network structures of GRNs.

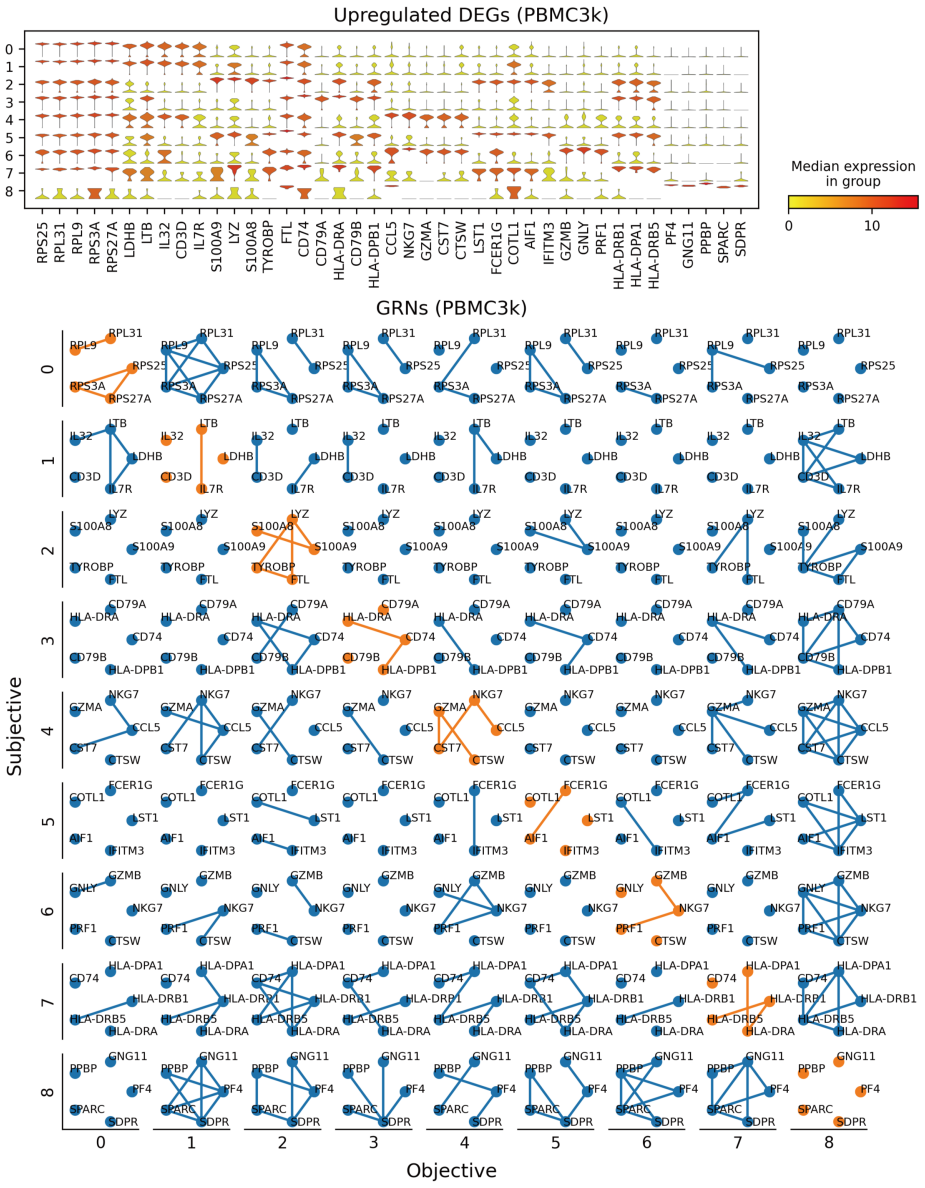

Figure 2: Gene expression patterns of clusters in PBMC3k