# Correlation Test of Two Single-Cell Data Sets with Manifold Distance Mantel Test

Yamada Ryo[1,*], Takahashi Tomohiro [1]

**1 Unit of Statistical Genetics, Center for Genomic Medicine, Gratuate School of Medicine, Kyoto University**

*** yamada.ryo.5u@kyoto-u.jp**

## Abstract

With the development of experimental technologies, a population of single cells are analyzed with multiple platforms at the same time. The measurement of the correlation among the different platform data sets is one of the fundamental approaches to integrate them. We propose a Mantel test [6] to measure correlation between two distance matrices, where distance between two cells is defined as graph distance in knn graphs that represent a manifold of the cell population in the feature space of each platform. The knn graph-based manifold representation of cell population in feature spaces has been accepted for single cell data analysis as a part of UMAP method [1]. The mantel test indicates statistical significance of correlation between two distance matrices when comparing the whole features of two platforms.

## Introduction

It is not easy to measure multiple omics profiles simultaneously of single cells yet, but the technology for it has been in progress [5]. Besides the conventional omics platforms, such as genome, epigenome, transcriptome and proteome, individual cells have various information, that are not the states of biomolecules, but cellular features, including location [2], morphology, mobility [**?**]. These non-molecular cellular features have been also recorded in single cellular fashion and they are ready to be analyzed with biomolecular omics data sets. We propose a method to test the overall correlation between two simultaneous feature sets of a population of single cells, e.g., mobiligy-morpholgy vs. transcriptome.

## Method

### 0.1 knn graph and manifold in feature space

The quantitative measurement of single cells for omics and other features are multivariate and in many cases their dimensions are very high. Although the dimensions are high, the cell population seems to take a relatively lower dimensional manifold and the cellular samples have been successfully visualized in lower dimensional space using non-linear dimension reduction methods, such as t-SNE and UMAP [3]. The latter method, UMAP, generates a connected graph based on knn-graph with the feature records to estimate the low-dimensional manifold and consider the dissimilarity between cells as their graph-distance or the estimate of distance along the manifold of cell population, subsequently project the connected graph in the target dimensional space with graph visualization algorithms [1]. Therefore, the matrix of pairwise distance of cells in the graph is reasonably representing the dissimilarity structure of sample cells.

When a set of sample cells are measured for two different feature sets, each feature records identifies a connected graph, respectively, and the graphs give two distance matrices whose rows and columns stand for the cells in the same order.

Although there is a freedom of ambiguity in the choice of value of k of knn graph, the question is shared with the UMAP method and we don't discuss further on it in this paper, but we assume we can construct a connected graph from each single cellular feature set and subsequently make a distance matrix based on the graph. The data records should be appropriately evaluated and checked for their quality before graph construction, as a part of regular quantitative biology/single cell experiment data analysis pipelines.

### 0.2 Mantel test for correlation of two distance matrices

The Mantel test is the statistical test for the correlation between two distance matrices. It is categorized in the non-parametric class and computes the significance of the correlation through permutations of the rows and columns of one of the input distance matrices. [6]

The permutations return the null distribution of correlation coefficients of two distance matrices and p-value for the particular data set pair.

## Application to a toy dada set pair

The codes in R language is available in the github https://github.com/ryamada22/atom/tree/master/overlaef/Mantel4scOmics/singleCellMantelTes , where a toy data set pair was randomly generated and the UMAP method was applied to the two data sets and the corresponding knn graphs were extracted. The graph distance matrices were calculated and they were tested with Mantel test.

## 1 Discussion

Once the correlation is found to be statistically significant overall, it is considered rational to investigate the contributing individual features of one platform to the whole other platform to further data-mine the two platform single-cell data sets. We applied our method to single cell data sets of mobile cells; in vivo cellular shape and movement were evaluated with the method proposed by Yusri et al., [4] and 26 shape/movement features were extracted. The microscopically observed cells were identified and taken out individually and their single cell transcriptome data were obtained. We applied our proposing method to these two feature data sets and identified weak but statistically significance was observed (Data not shown for further detailed biological investigation).

## Acknowledgments

Funding infromation : Grant numbers JPMJCR1502 and JPMJCR15G1. Core Research for Evolutional Science and Technology (CREST) URL of each funder website: https://www.jst.go.jp/kisoken/crest/en/. DO, grant number JP19J14816. KAKENHI Grant-in-Aid URL of each funder website: https://www.jsps.go.jp/english/e-grants/

## References

1. E. Becht, L. McInnes, J. Healy, C. A. Dutertre, I. W. H. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol*, Dec 2018.

2. D. J. Burgess. Spatial transcriptomics coming of age. *Nat Rev Genet*, 20(6):317, 06 2019.

3. C. N. Heiser and K. S. Lau. A Quantitative Framework for Evaluating Single-Cell Data Structure Preservation by Dimensionality Reduction Techniques. *Cell Rep*, 31(5):107576, 05 2020.

4. Y. D. Heryanto, C.-Y. Cheng, Y. Uchida, K. Mimura, M. Ishii, and R. Yamada. Integrated analysis of cell shape and movement in moving frame. *bioRxiv*, 2020.

5. J. Lee, D. Y. Hyeon, and D. Hwang. Single-cell multiomics: technologies and data analysis methods. *Exp Mol Med*, 52(9):1428–1442, 09 2020.

6. N. Mantel. The detection of disease clustering and a generalized regression approach. *Cancer Res*, 27(2):209–220, Feb 1967.