# Variability of Single Cell Gene Expressions for Different Cell Types

**Ke He, with supervisor Heejung Shim**
School of Mathematics and Statistics, The University of Melbourne
khhe1@student.unimelb.edu.au.com

## Introduction

Individual cells from the same cell type can have different gene expressions. Such **transcript-level** variations potentially lead to **protein-level** differences and are found to be related to certain biological mechanisms. For example, gene variability in cancer are found to be correlated with its proficiency at metastatic colonization and resistance to chemotherapy [1].

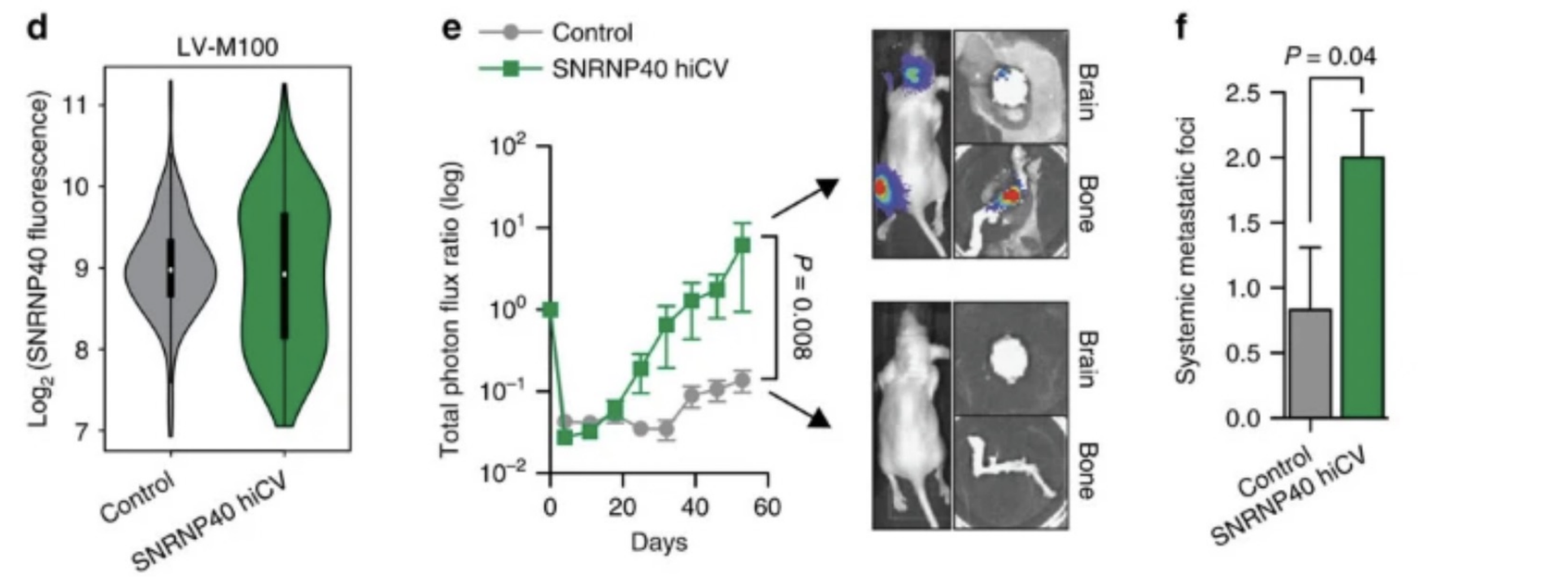


**Figure 1:** High expression variability of spliceosomal gene SNRNP40 leads to increased metastatic capacity [1]

While there are many existing researches on measuring the variability of particular genes, there is no current research focusing on measuring the overall variability of gene expressions for different cell types. Here, we built a simple two-step variance based method to tackle this problem. We also developed two metrics using the method and tested their performances on a real data set.

## The Method

### Step 1: Dimension Reduction

In single cell data analysis, the data matrix is usually **sparse and high dimensional**. Hence, we will first apply dimension reduction and choose factors that best summarize the variability. **In this preliminary analysis, we will use PCA.**

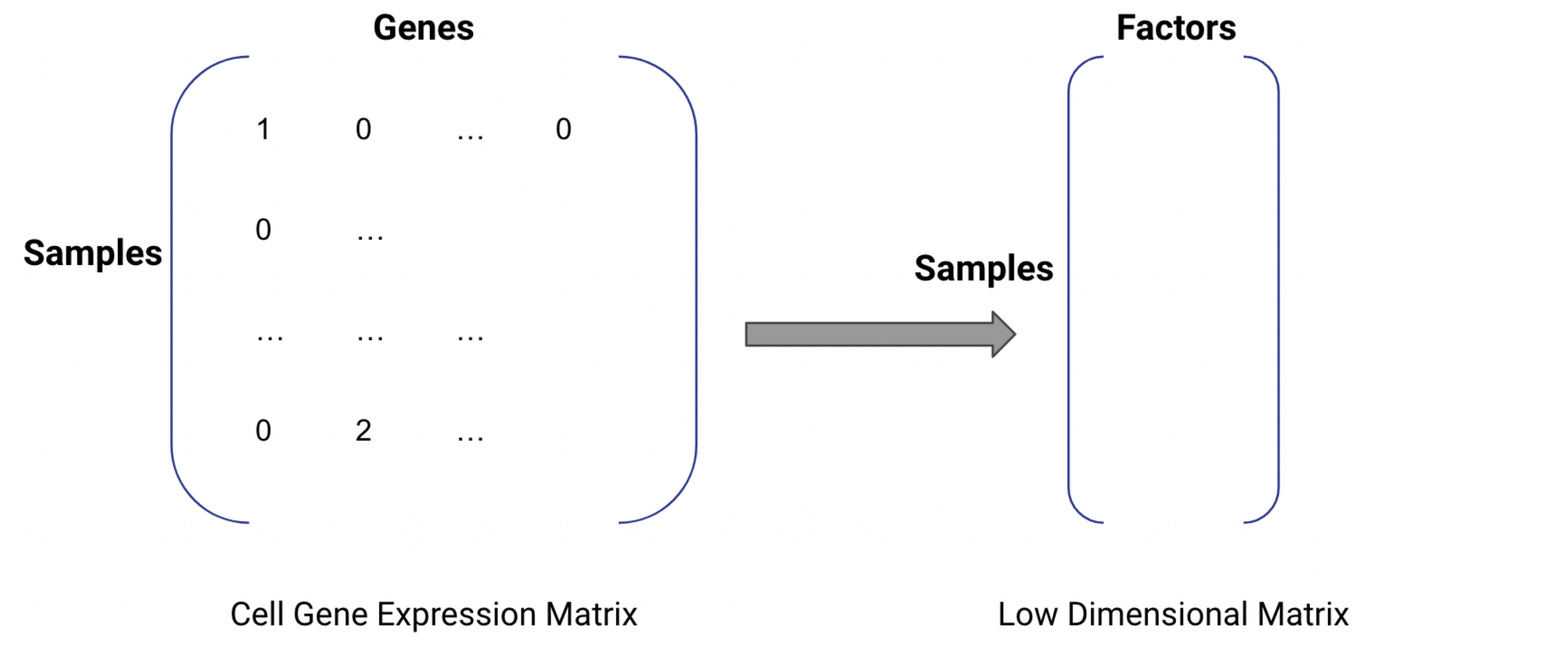


**Figure 2:** Dimension reduction

### Step 2: Variance-Covariance Matrix

To measure the variability for a cell type, we first calculate the variance-covariance matrix of the cell type. Then we derive a metric to summarize the information in the variance-covariance matrix. A naive way to do this is by just looking at the diagonal entries. The diagonal entry $V_i$ is the cell type's variance with respect to the $i^{th}$ factor. We proposed two metrics to summarize the diagonal entries:

- **Sum** = $\sum_{i=1}^{n} V_i$
- **Product** = $\prod_{i=1}^{n} V_i$

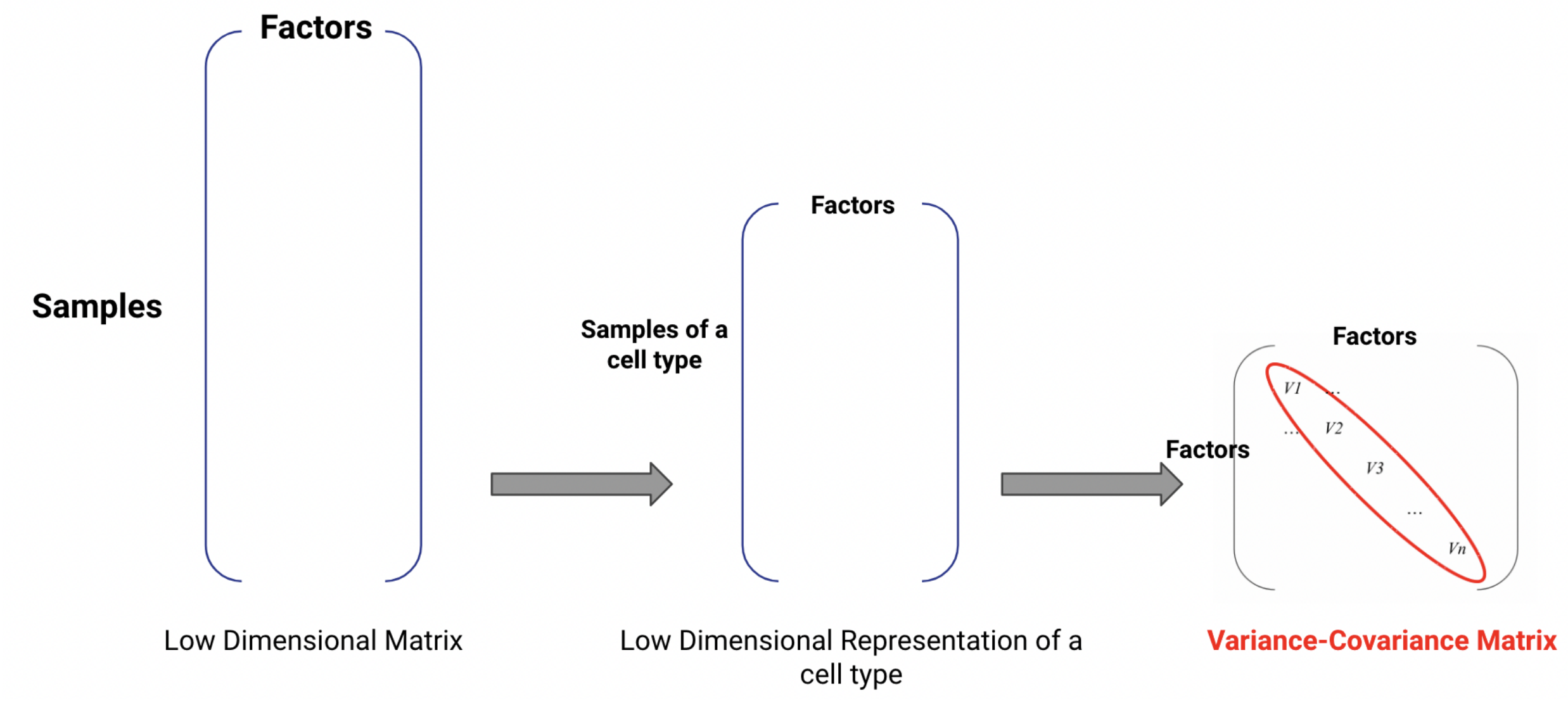


**Figure 3:** Derive the variance-covariance matrix for a cell type

## Data Set

We tested the metrics on transcriptional and cellular diversity of the human heart data [2]. The data set consists of 287,269 labelled samples of 17 cell types and covers 33,694 genes.

## Results and Findings

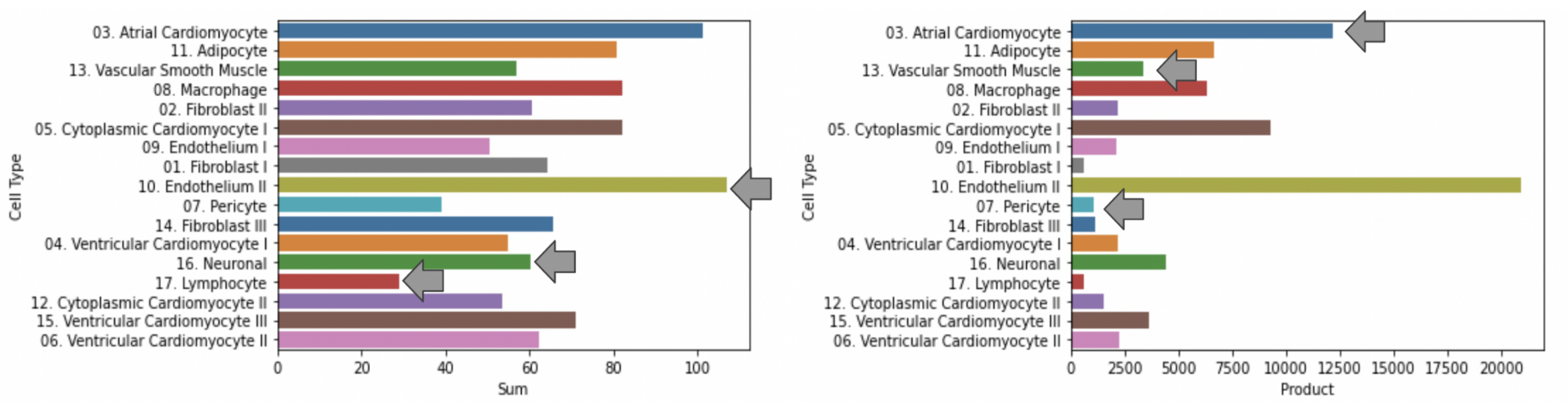


**Figure 4:** Sum and product results

We chose one group of cell types (marked by grey arrows in Figure 4) to visualize in the pair-wise PC (principle component) space for each metric we have. Each group contains three cell types whose variability are identified as high, medium and low respectively.

### Sum

10.Endothelium II: 107.0 16. Neuronal: 60.1 17. Lymphocyte: 29.0

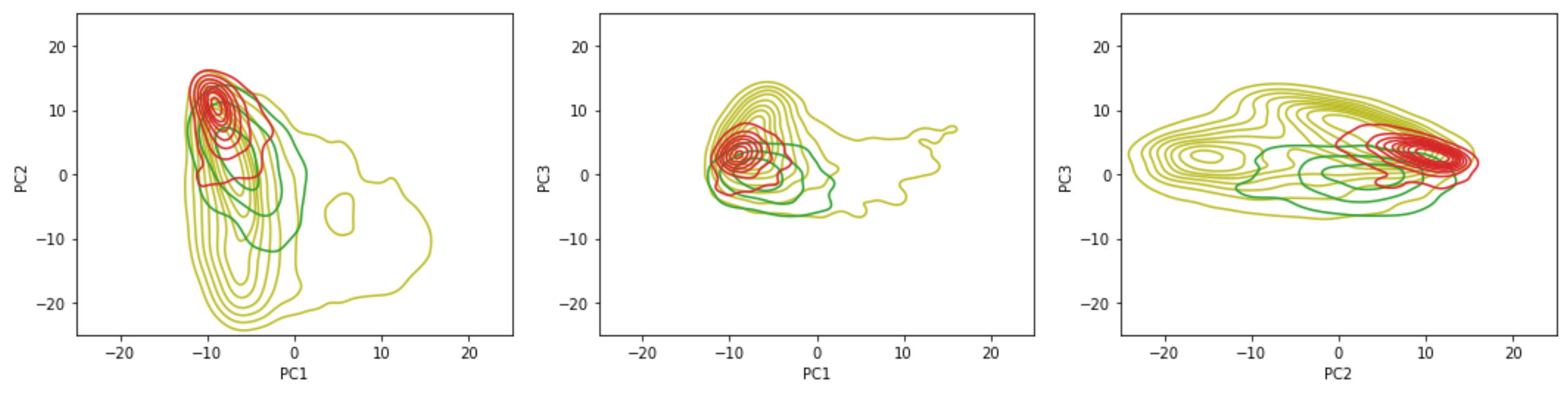


**Figure 5:** Cell types with high, medium and low variability determined by sum

### Product

03. Atrial Cardiomyocyte: 12165.1 13. Vascular Smooth Muscle: 3364.1 07. Pericyte: 1086.8

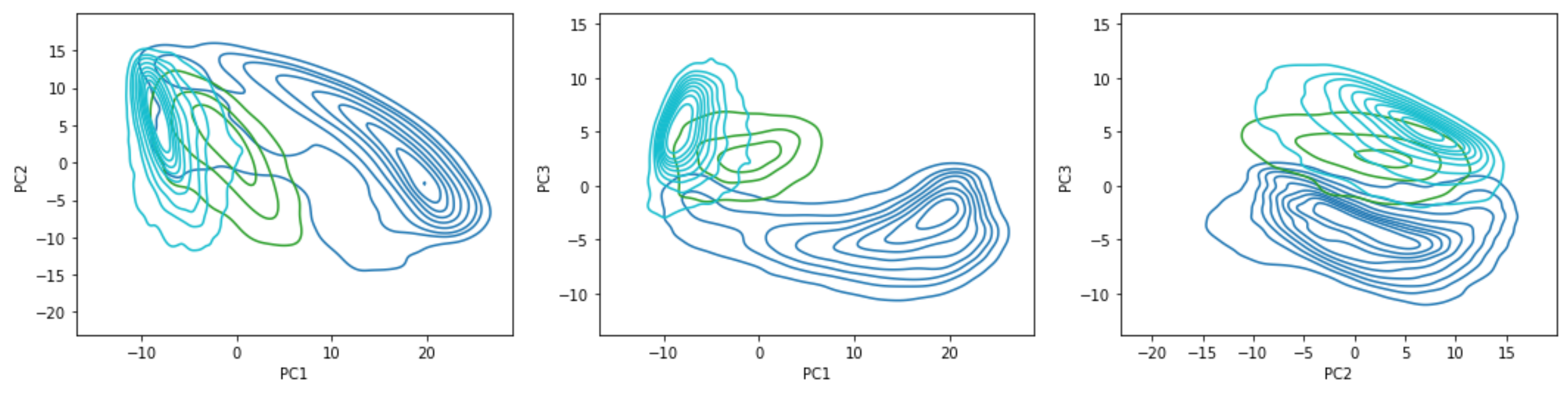
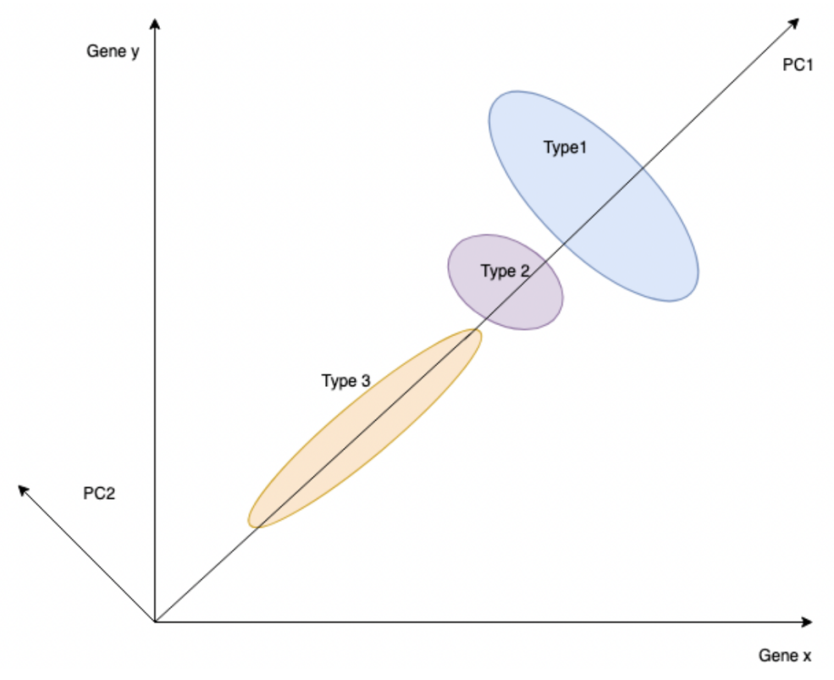


**Figure 6:** Cell types with high, medium and low variability determined by product

### Potential Problem with Product



| | PC1 | PC2 | Product |
|---|---|---|---|
| Type 3 | 3 | 0.3 | 0.9 |
| Type 2 | 1 | 0.9 | 0.9 |

**Figure 7:** A theoretical example of a potential problem with the product

One potential issue with the product is that instead of looking at absolute values, the product actually compares ratios. For example, in Figure 7, ***Type 3*** is more variable than ***Type 2*** by visual observation. However, they will be measured as the same using the product. More examples can be found in the Github repository[3].

## Conclusions

In this preliminary analysis, we found that both the sum and the product capture the variability of single cell gene expressions for different cell types in real data. This shows the potential of the two-step method we proposed.

## Forthcoming Research

- **Integrate covariances**: It is desirable to integrate the off-diagonal entries in the variance-covariance matrix and penalize cell types which show high correlations between the factors.
- **Alternative dimension reduction methods**: Other linear methods such as factor models and matrix factorization, as well as non-linear ones like variational autoencoder could be experimented.

## Reference

[1] Nguyen, A., Yoshida, M., Goodarzi, H., & Tavazoie, S. F. (2016). Highly variable cancer subpopulations that exhibit enhanced transcriptome variability and metastatic fitness. Nature communications, 7, 11246. https://doi.org/10.1038/ncomms11246

[2] Tucker, N.R., Chaffin, M., Fleming, S.J., Hall, A.W., Parsons, V.A., Akkad, A.D., Herndon, C.N., Arduini, A., Papangeli, I., Roselli, C., Aguet, F., Choi, S.H., Ardlie, K.G., Babadi, M., Margulies, K.B., Stegmann, C.M. & Ellinor, P.T. (n.d.). *Transcriptional and Cellular Diversity of the Human Heart*. https://singlecell.broadinstitute.org/single_cell/study/SCP498/transcriptional-and-cellular-diversity-of-the-human-heart

[3] Github repository link: https://github.com/KeHe01/CellTypeVariability.git