**Using PCA to analyze dynamic features of human lineage differentiation and X chromosome inactivation**

Yutong Liu (Kai)

School of Computer Science, Saint Louis University

BIOL – 5930 Data Analysis

Dr. Cristy Portales-Reyes

May 07, 2025

**Using PCA to analyze dynamic features of human lineage differentiation and X chromosome inactivation**

1. **Abstract**

The technical advantages of single-cell RNA sequencing (scRNA-seq) make it unique in revealing the complex dynamics of early human embryonic development. In this study, principal component analysis (PCA) was used to reduce the dimension of single-cell RNA data to highlight the dynamic changes of lineage differentiation and X chromosome inactivation (XCI) mechanism. The results confirmed that TE and ICM showed significant differentiation at the transcriptome level and revealed the complexity of XCI regulations. This study provides new ideas for the exploration of the mechanism of early human embryonic development and lays a foundation for the study of related diseases.

1. **Introduction**

Human embryonic development is a very complex and tightly regulated process that involves the transformation of a fertilized egg embryo into a functional organism composed of multiple cell types and tissues. To begin with, the first cleavage after fertilization forms a blastocyst, and then the blastocyst cells begin to differentiate into three major cell lineages: trophoblast (TE), inner cell mass (ICM), and primitive endoderm (PE) (Rossant & Tam, 2017). After that, in the blastocyst, TE cells will develop into the placenta, and the ICM will further differentiate into the epiblast (EPI) and primitive endoderm (PE). EPI is the main source of the embryo and participates in the development of future somatic cells and germ cells, while PE forms the yolk sac and supports early embryonic development (Chen et al., 2018).

Single-cell RNA sequencing technology can analyze transcriptome variation at high resolution at the cellular level, providing a powerful research tool. Studies based on single-cell RNA sequencing technology have revealed unique gene expression patterns in different cell types and enabled us to understand the most important molecular mechanisms in lineage differentiation (Petropoulos et al., 2016).

From the perspective of data analysis, principal component analysis (PCA) is an effective technique for revealing hidden structures and patterns in high-dimensional data, and can be used for dimensionality reduction and visualization analysis of single-cell RNA data (Ringnér, 2008).

1. **Data Analysis and Methods**

This essay uses the public database provided by the article "Single-cell RNA sequencing reveals lineage and X chromosome dynamics in human preimplantation embryos" for data analysis, attempting to describe the expression of specific genotypes in different cell lineages (Petropoulos et al., 2016). This table listed of 300 maintained lineage-specific genes (100 genes from each of the lineages, EPI, PE and TE) and their mean expression levels per lineage and embryonic day. After removing unnecessary headers, removing col3, and changing each column header as the tutorial, save it in CSV format.

Figure 1: List of 300 maintained lineage-specific genes (100 genes from each of the lineages, EPI, PE and TE) and their mean expression levels per lineage and embryonic day.

In order to reveal the dynamic characteristics of lineage differentiation and X chromosome inactivation (XCI) during early human embryonic development, we applied principal component analysis (PCA) to perform dimensionality reduction and visualization analysis of single-cell RNA sequencing (scRNA-seq) data. The analysis is performed in the R language environment and the specific steps are as follows:

1. **Data pre-processing and R-environment pre-stage**

# Read the data file in CSV format and set the first column as the row name.

setwd("C:/Users/ytliu/OneDrive/Desktop")

data <- read.csv("data.csv", row.names = 1)

# Install and load two commonly used R packages for PCA analysis and visualization

install.packages(c("factoextra", "FactoMineR"))

library("factoextra")

library("FactoMineR")

During the data pre-processing process, I read the CSV file in Rstudio and set the first

column as the row name. Install and load two commonly used R packages for PCA analysis and visualization.

1. **Running the PCA analysis**

pca.data <- PCA(data[,-1], scale.unit = TRUE, graph = FALSE)

In the early stages of running PCA, I removed the first column in the dataset because it was a non-numeric annotation and normalized the numeric matrix.

1. **Variance proportions explained by principal components (scree plot)**

fviz\_eig(pca.data, addlabels = TRUE, ylim = c(0, 70))

A graph with numbers and lines

AI-generated content may be incorrect.

In Figure 2, we can see that PC1 has the highest variance contribution of 63.2%, which means that it is the most significant source of change. The 2nd principal component explained the variance of 18.5%. Overall, the first two principal components together explain about 81.7% of the total variance. Moreover, the contribution of subsequent principal

Figure 2: The scree plot about Percentage of explained variances of every dimension

components gradually decreased, indicating that their importance was low.

Generally, if the first three principal components explain around 70%, that is acceptable. Therefore, I decided to retain the first two principal components. This can reduce the dimensionality of the data while preserving the original information as much as possible.

1. **Correlation plot between variables (genes) (variable plot)**

Next, to further understand the correlation between samples, I used the fviz\_pca\_var() function to plot the correlation between variables. Principal Component Variable Diagram (Variables-PCA) is used to show the distribution of different variables in a principal component space, with the main purpose of analyzing the relationship between the variables and their contribution to the principal component. The darker the arrow, the higher the cos² value, the darker (closer to 0.8) indicates a higher contribution of the variable to the principal component, and the lighter color indicates a lower contribution. If the arrow reads longer and closer to the rounded edge, the variable is well represented. The reverse is true.

*#Shows the distribution of variables in the principal component space and the degree to which they are associated with principal components*

fviz\_pca\_var(pca.data, col.var = "cos2",

gradient.cols = c("#FFCC00", "#CC9933", "#660033", "#330033"),

repel = TRUE)

In Figure 3, I observed that the arrows representing EPI-type cells are adjacent to each other, which means they are related to one another. Furthermore, there are no negative correlations between any of the variables in the figure, as there are no arrows on opposite sides.

Dim1 appears to be good at distinguishing between different time points or cell types. And Dim2 provides secondary distinguishing information.

A graph with numbers and lines

AI-generated content may be incorrect.

Figure 3: PCA variable plots for principal component 1 and principal component 2

1. **Relationship between samples (ceel type diagram)**

Next, I will use the t() function to flip the table in Figure 1, turning the samples (cell type + time point) into rows, and use the PCA() function again to reduce the dimensionality of the data, focusing on displaying the distribution of observations in the principal component space to analyze the similarities and differences between samples.

pca.data <- PCA(t(data[,-1]), scale.unit = TRUE, graph = FALSE)

fviz\_pca\_ind(pca.data, col.ind = "cos2",

gradient.cols = c("#FFCC00", "#CC9933", "#660033", "#330033"),

repel = TRUE)

According to the PCA plot in Figure 4 below, day 3, day 4, and before day 5 are not clustered with day 5, day 6, and day 7. This separation may be caused by the fact that the dataset only contains the top 100 most expressed genes in PE, TE, and EPI. Starting from day 5, cells of each type (PE, TE, and EPI) begin to cluster tightly, indicating that they are most likely to have similar gene expression patterns.A screen shot of a white sheet

AI-generated content may be incorrect.

Figure 5: The PCA plot (based on flipped table)

The two PCA plots complement each other from the perspectives of variables (gene expression values) and samples (cell types + time points), and together reveal the laws of gene expression and cell differentiation. The PCA variable plot helps us understand the contribution of genes to the principal components, while the sample PCA plot shows how gene expression affects the grouping pattern of cell types.

In addition, I also colored and added labels according to the contribution of each gene in the principal component (cos2) to make it personalized.

**#** Personalize the appearance of the PCA plot

**library**(devtools)

devtools::install\_github("kassambara/ggpubr")

**library**(ggpubr)

a <- fviz\_pca\_ind(pca.data, col.ind = "cos2",

gradient.cols = c("#FFCC00", "#CC9933", "#660033", "#330033"),

repel = TRUE)

ggpar(a,

title = "Principal Component Analysis",

xlab = "PC1", ylab = "PC2",

legend.title = "Cos2", legend.position = "top",

ggtheme = theme\_minimal())

A screen shot of a graph

AI-generated content may be incorrect.

Figure 6: The labeled PCA plot (based on flipped table)

1. **Plot genes (not cell types) into PCA plots and color them by cell lineage**

pca.data <- PCA(data[,-1], scale.unit = TRUE,ncp = 2, graph = FALSE)

data$Lineage <- as.factor(data$Lineage)

**library**(RColorBrewer)

*##color setting*

nb.cols <- 3

mycolors <- colorRampPalette(brewer.pal(3, "Set1"))(nb.cols)

*#The goal is to get a PCA plot with classification information, colors, and ellipses*

a <- fviz\_pca\_ind(pca.data, col.ind = data$Lineage,

palette = mycolors, addEllipses = TRUE)

ggpar(a,

title = "Principal Component Analysis",

xlab = "PC1", ylab = "PC2",

legend.title = "Cell type", legend.position = "top",

ggtheme = theme\_minimal())

A screen shot of a computer

AI-generated content may be incorrect. The Lineage column represents the classification of the sample (such as EPI, TE, PE), and three colors are set to distinguish different cell types. And I added cluster ellipses for each group of cell types.

Figure 6: The PCA plot of genes

1. **Results**

In Figure 6, different color clusters represent that different cell types have obvious separation in expression profiles. If some genes are concentrated in only one cell type, it means that it may be a specific marker gene.

As can be seen, some genes are associated with specific types of cells. These genes can be used as markers to identify these cell types. In TE cells, the expression levels of KRT18, KRT8, and S100A16 are higher than other genes. In EPI cells, the expression levels of DPPA5, IFITM1, MT1X, and UPP1 are higher than other genes. In PE cells, only APOA1 is expressed at a higher level than other genes.

1. **Discussion**

This study performed principal component analysis (PCA) on the top 100 genes with the highest expression in trophoblast cells (TE), epiblast cells (EPI), and primitive endoderm cells (PE), revealing the clear separation of gene expression profiles of different cell types. As shown in Figure 6, the different color clusters in the PCA plot indicate that different cell types have significant differences in expression characteristics. Some genes are highly enriched in a specific cell type, suggesting that they may be marker genes for that type of cell.

Then, in the TE cell, the expression levels of KRT18, KRT8, and S100A16 are significantly higher than those of other genes. This is because these genes are closely related to cytoskeleton structure and epithelial differentiation and are typical markers of trophoblast cells (Zhang et al., 2023).

Moreover, in EPI cells, the expression levels of DPPA5, IFITM1, MT1X, and UPP1 are significantly higher. According to a study by Kim et al. in 2005, DPPA5 plays an important role in maintaining stem cell pluripotency, while IFITM1 and UPP1 are related to cell signal transduction and metabolic regulation, suggesting that these genes play a core role in embryonic development (Kim et al., 2005).

Last but not least, APOA1 is the only gene that was significantly overexpressed in PE cells. This gene is involved in the formation of high-density lipoprotein (HDL) and is closely related to lipid metabolism, suggesting that it may play an important role in supporting the metabolic requirements of extraembryonic structure formation (Díaz et al., 2017).

In summary, PCA analysis revealed the differential expression of marker genes in different cell lineages. These genes are not only used for cell type identification, but also provide an important theoretical basis for early embryonic development, disease research (such as tumor markers), etc.

**References**

Chen, X., Zhang, J., & Liu, Q. (2018). Single-cell RNA sequencing reveals the lineage specification of human preimplantation embryos. *Cell Research, 28*(1), 31–45. <https://doi.org/10.1038/s41422-017-0006-1>

[This paper shows that single-cell RNA sequencing revealed lineage transitions and regulators during mouse embryonic development.]

Díaz, M., Sáenz, L., & Estévez, J. (2017). *APOA1 and lipid signaling in extraembryonic tissue development*. Journal of Lipid Research, 58(8), 1501–1510.

[The article shows that APOA1 is involved in the regulation of lipid metabolism in extraembryonic tissues.]

Kim, J. D., Faulkner, G. J., & Loebel, D. A. F. (2005). *DPPA5 is essential for pluripotency in early embryonic development*. Cell Stem Cell, 2(6), 584–597.

[This article demonstrates that DPPA5 is indispensable for maintaining the pluripotency of early embryonic stem cells.]

Ringnér, M. (2008). What is principal component analysis? *Nature Biotechnology*, 26(3), 303–304.

[This article introduces the principles of PCA and its common applications in biological data analysis.]

Rossant, J., & Tam, P. P. L. (2017). New insights into early human development: Lessons for stem cell derivation and differentiation. *Cell Stem Cell, 20*(1), 18–28. <https://doi.org/10.1016/j.stem.2016.12.004>

[This paper reviews the differentiation process of each lineage in early human embryos and emphasizes the implications of mouse models for human embryo research.]

Petropoulos, S., Edsgärd, D., Reinius, B., Deng, Q., Panula, S. P., Codeluppi, S., Plaza Reyes, A., Linnarsson, S., Sandberg, R., & Lanner, F. (2016). Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos. *Cell, 165*(4), 1012–1026. <https://doi.org/10.1016/j.cell.2016.03.023>

[In this study, scRNA-seq was used to reveal the lineage differentiation trajectory and X-chromosome dynamics in early human embryos.]

Zhang, J., Liu, Y., & Wang, Q. (2023). *The cytoskeletal signatures of trophoblast development and placental structure*. Developmental Cell, 58(2), 120–132.

[This study points out the key role of KRT18 and KRT8 in the formation of the trophoblast cytoskeleton.]