NANYANG TECHNOLOGICAL UNIVERSITY

Story-telling of Hepatocellular Carcinoma Single Cell genes with Graphics and Visualization

(BS6203: STORY-TELLING WITH GRAPHICS AND VISUALIZTION)

GROUP 4

ZHANG ZHU

YE SHUCHEN

HAN WENHAO

LI GUANCHEN

Content

Introduction

Background

Hepatocellular carcinoma (El-Serag & Mittal, 2013) is the main subtype of liver cancer, which leads to high mortality worldwide. One of the obstacles to combating it is hematogenous dissemination, namely the major route of HCC metastasis.

Data

We used single-cell RNA sequencing to isolate 113 CTCs from four key vascular sites along the dissemination pathway in 10 HCC patients, identifying transcriptional heterogeneity within the circulatory system and between vessels.

Audience

Our main readers are biotechnology scientists, and we need to point them to genes with greater variation in four key vascular sites of transmission in HCC patients. So our task is to present our modeling process in a simpler and easier-to-understand visualization so that they can understand and trust our results.

Significance

In a report, in addition to data analysis and modeling, how to present the information and results we get to the reader is also a very important part. Since readers of our paper are not necessarily familiar with biological information or mathematical models, the necessary images and tables can help them understand our models and results.

But is any image suitable for a paper? According to Gestalt psychology, our visual properties of the human eye and the brain's processing behavior of graphics are worth studying. Accurate and appropriate images can help us understand, but wrong images can lead to confusion or misunderstandings for readers, which is not conducive to the transmission of information. Therefore, how to accurately use images to convey the correct information is the purpose of our paper.

Display Data

Sample grouping distribution

Before modeling, we need to understand the data and present it so that we can do proper data preprocessing. The first step we need to do is to show the distribution of the data. Since we are using multidimensional data, we can use a boxplot for this step.

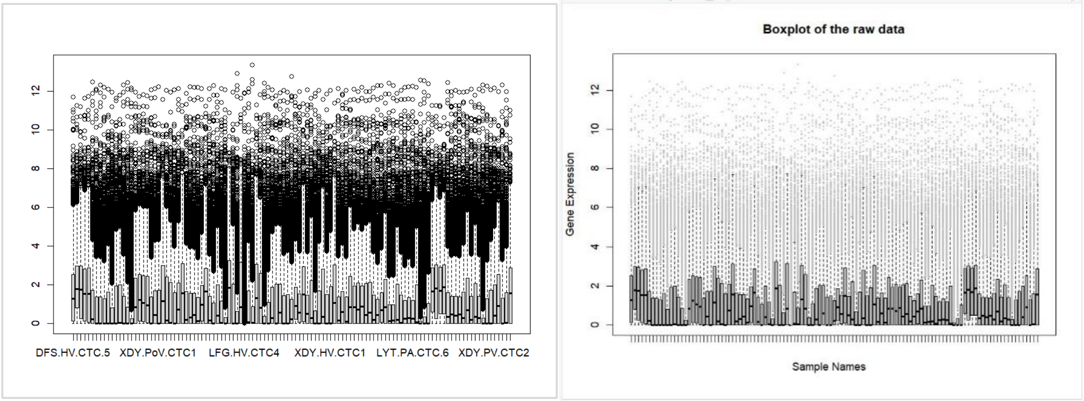

Fig 1: Boxplot of the raw data

Before drawing a graph, we all have a goal, and in this graph, our goal is to see the distribution of the data.

Make it clear what's important: In the initial boxplot, however, the outliers stack up quite dramatically due to their hollow circle shape. So, we need to reduce the presence of outliers. This can be achieved by changing the color of the outliers to a low-saturated gray and changing the shape to a small solid circle.

Explain the data information clearly: In the original data, we did not indicate the title, x-axis, and y-axis. Therefore, people who see this picture for the first time will be very confused and not understand what this picture is trying to show. Therefore, we clearly marked the title and the labels of the x-axis and y-axis in the improved boxplot.

To avoid confusion: In the initial figure, the sample names on the x-axis cannot be displayed all, so the system randomly displays several sample names. However, this can be confusing to the reader, because not only do we not know which samples are named, but we also wonder if these are the most important samples. So, no information is better than confusing information, we choose to remove all sample names in the improved boxplot.

Overall distribution

Looking at the distribution of different samples in the data, we are also curious about the overall distribution. Therefore, we can draw a histogram to show it.

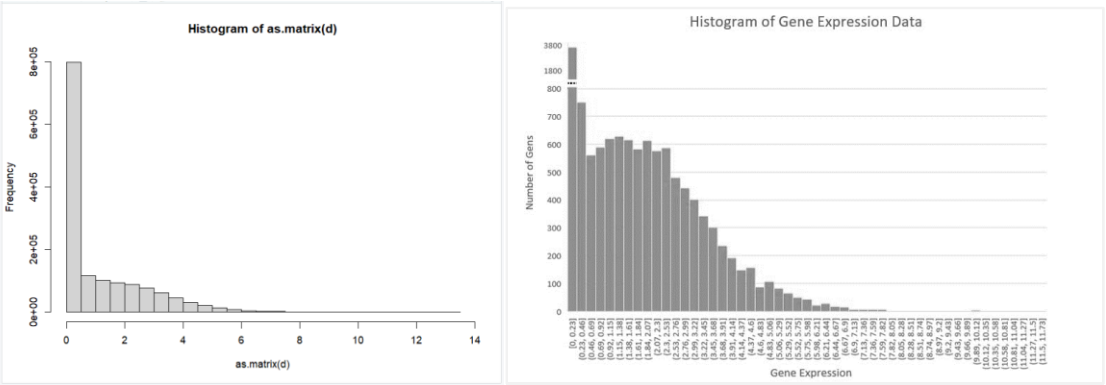

Fig 2: Histogram of the raw data

Avoid the influence of extreme values: We want to observe the overall distribution, but the existence of extreme values may prevent us from observing the distribution of small data well. So in this case we can use the break method to truncate the extreme values, to shorten outlier elements that would otherwise compress the dynamic range of the data.

Auxiliary line: Since the right column is far away from the y-axis, there is no way to judge its value. At this time, adding a faint auxiliary line to the background can help the reader to judge the size of the data.

Labels for the x-axis and y-axis: In most cases, when the x-axis names are very long, the author will choose to tilt the names at a forty-five-degree angle or interchange the x-axis and y-axis for readability. In this project, however, our focus is on the number of genes with different gene expression levels, not gene expression levels. So, we still use the quantity as the y-axis and don't rotate the x-axis names for a better view.

Data information: As with the first set of comparison charts, we need information about the SD data. Such as the title of the chart, and the labels of the x-axis and y-axis.

Significance of Chart Improvements:

In the original image, we can only see that the data very close to 0 has an exceptionally large frequency, and we do not have a good understanding of the distribution of other data. However, when the data is improved, we can see that the vast majority of the data is at the level of 0-0.23. The data are not monotonically decreasing, and there is a slight increase in the data of gene expression levels between 0.96 and 1.15. In addition, although the maximum gene expression amount is 11.73, data greater than 7 are very few.

Therefore, after the image is improved, we can read more information, and we also know that the data distribution is severely right-skewed. If we want to better use the data for modeling, we need to do some processing on the data, such as normalization and so on.

Distribution of data in different modeling groups

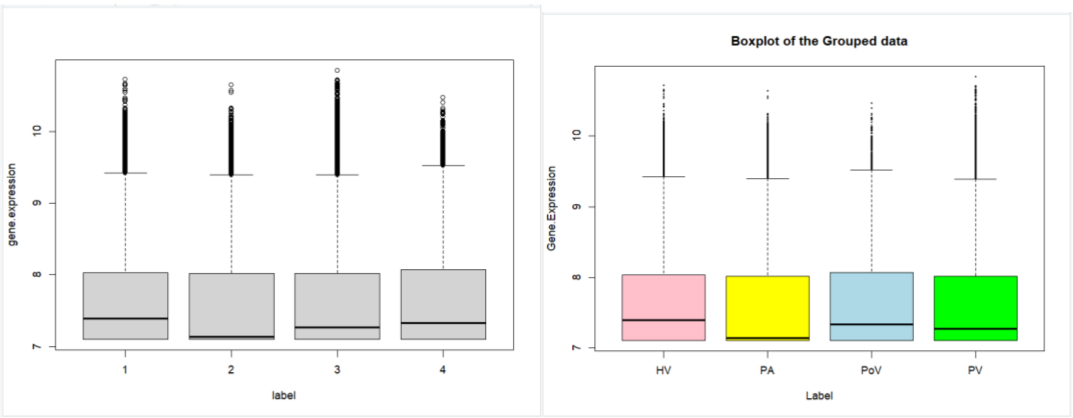

Fig 3: Boxplot of the grouped data

Clear grouping: All our samples can be divided into four parts, 'HV', 'PA', 'PoV', and 'PV'. In the initial boxplot, we did not indicate what the four groups were, only 1, 2, 3, and 4 instead. This can confuse the reader and is not conducive to a clearer perception of the data. Therefore, it is very important to make the data clear and readable.

Appropriate use of color: This is the first time we use color in this paper, because we draw pictures with the idea that color is for better reading, not just to make the image more beautiful. In addition, we choose four colors to represent the grouping. In order to better distinguish them, we choose four hues, red, yellow, blue, and green, instead of using different saturation of the same hue. Properly adjust their lightness to get better sensory effects.

Differential Expression Analysis

Before DEG

Before we do gene expression analysis (Zhang, Ou, & Zhang, 2004), we would like to see the percentage of genes in each location. We will use a donut chart which is an alternative plot of the pie chart to visualize it.

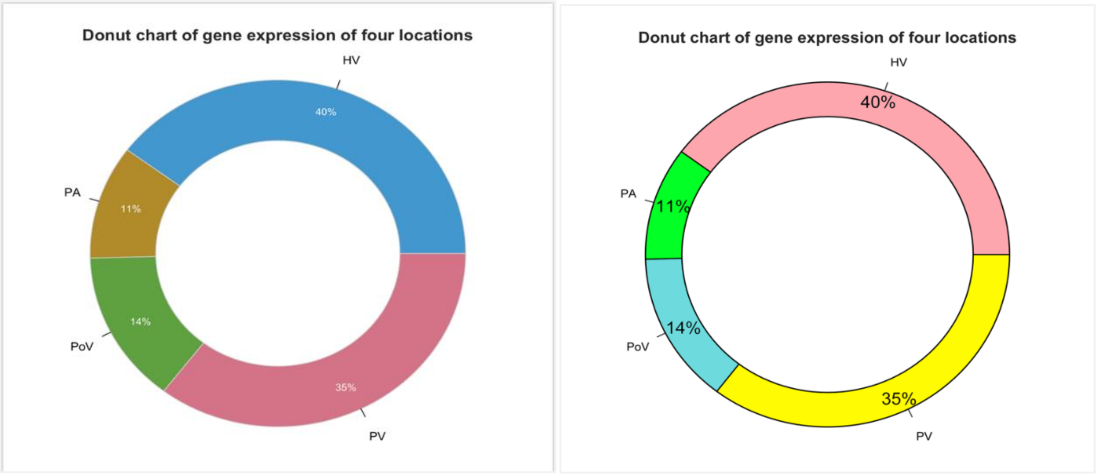

Fig 4: Donut chart of four locations

In the default plot, the errors may affect the clarity of reading the chart are the colors are different from we used to express these four locations, the font size of percentage is too small, the divided line between each part is not clear, and the color of percentage is white which may not clear to read under high saturation color background.

Then, we improved the donut chart. This time, we use the same color as previous plots, increase the font size of percentage, change the color of percentage to black, change the color of border line in to black and increase the width of border line a little to help the reader have a better distinguish of each part. Also, we changed the size of hole of be larger to make the donut chart be more aesthetic.

The violin plot is a hybrid of a box plot and a density plot, which shows peaks in the data. It is used to visualize the distribution of numerical data. Unlike a box plot that can only show summary statistics, violin plots depict summary statistics and the density of each variable.

We will plot the violin plot to see the density and distribution of the gene expression data after normalization in four different locations before doing further analysis.

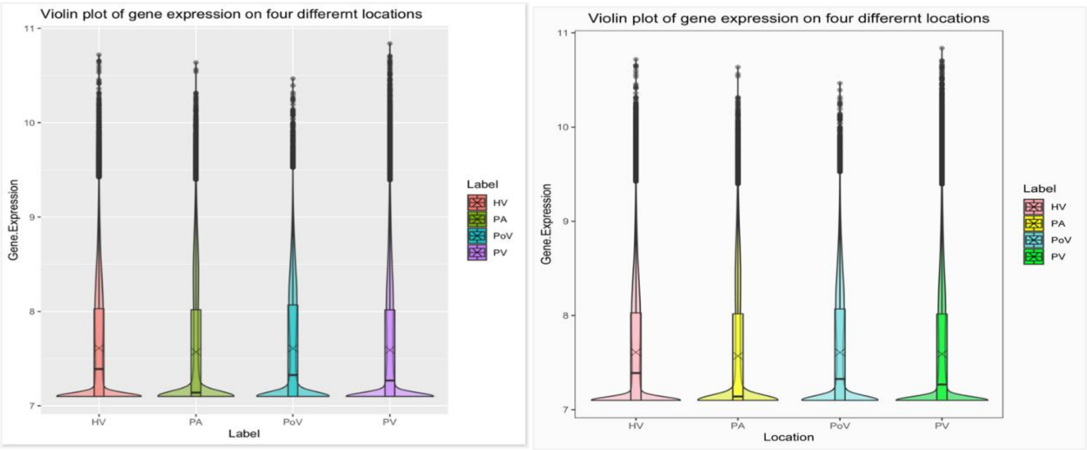

Fig 5: Violin Plot of four locations

For the default violin plot, the background and axis may let us hard to see rea the plot, the x-label is not correct, the title is not centered and the most important is the color, it used the default color and different from the color we used for these four locations in other plots.

In order to not make the reader confused, we improved our violin plot. In our improved plot, we get rid of the background and axis, the x-label is changed to 'Location', the title was centered and the color of all colors including the legend are changed to the same as the color we use in previous plots.

DEG Display

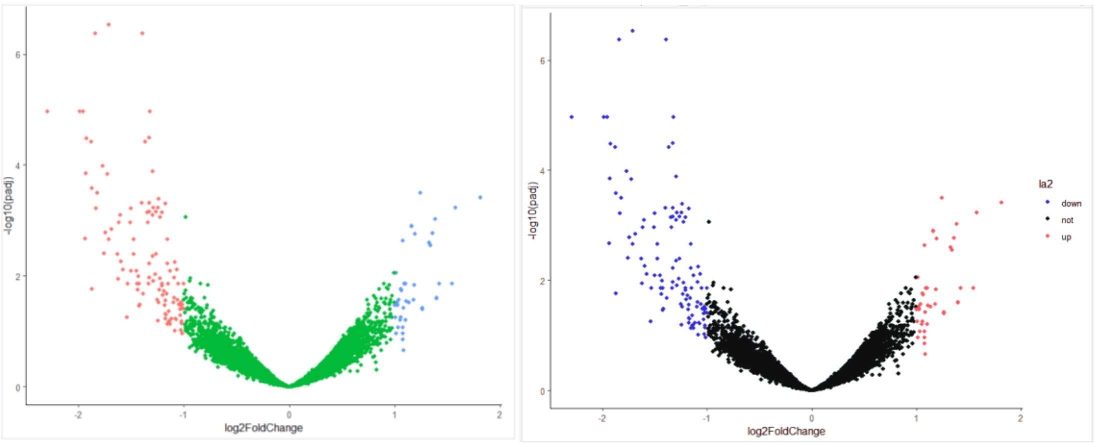

Fig 6: DEG volcano plot

From gene expression analysis we can draw volcano plots, however in the first plot we are only color-coded according to different data types. However, the classification details are not marked, which may confuse the reader. Besides, because the images use very bright colors, it is difficult to tell which data we are focusing on. Therefore, in the image on the right, we changed the color of the unimportant area in the middle to dark, and the rising and falling data on the left and right sides were changed to be the same as the other images in the article, and the labels were clearly listed on the right.

After DEG

After we used four methods to detect the differentially expressed genes, we would like to see the number of intersections among them. We can plot the Venn diagram to virtualize them.

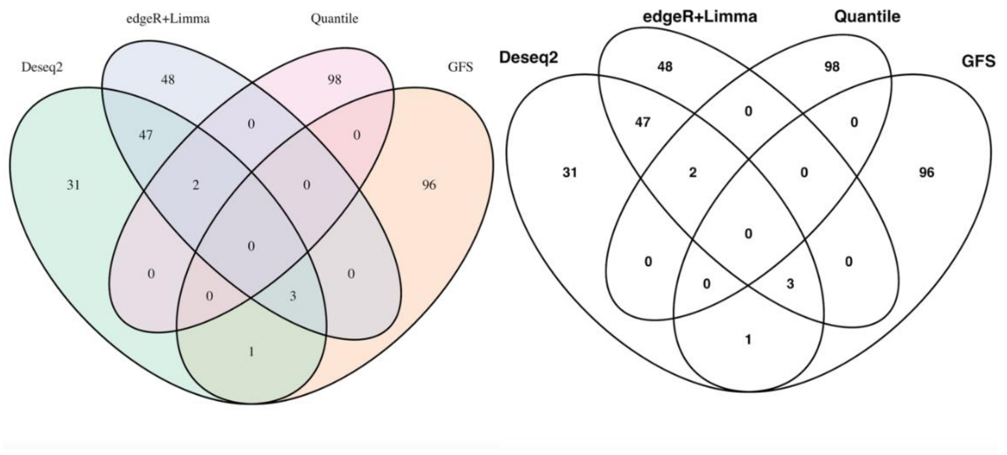

Fig 7: Venn of four DEG results

The first Venn diagram we plot has default colors and this may confuse the reader, but we also already used many different colors in our previous plots. In order of reduce the chance of misleading, we decide to not use any colors for the Venn diagram, and for the default plot, the font size was a little bit small and typeface thickness was too thick. Therefore, we increased the font size and thickness of font.

Machine Learning Task

Boxplot for feature distribution

Before we perform machine learning (Upstill-Goddard, Eccles, Fliege, & Collins, 2013) for important gene selection, we check each gene's expression distribution after normalization. The boxplot before improvement is shown below, we can hardly distinguish each box, the outliers overlap with each other and the x-axis is a mess.

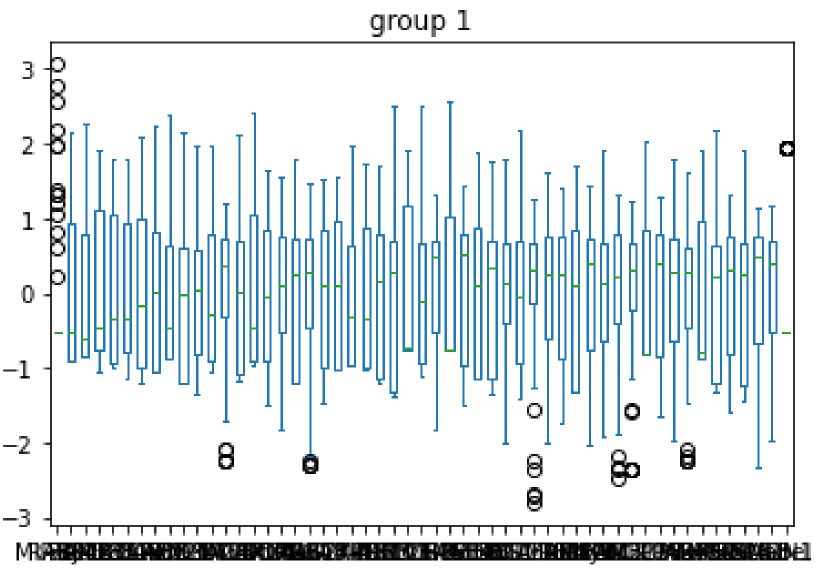

Fig 8: Initial Boxplot of group 1

The boxplot below is the improved one. We resized the figure so that each box has enough space and the comparison is easy to carry out. We rotate the x-tick value to handle the mess. We resized the outlier to handle the overlap and fill the box with different hues to make them easy to distinguish.

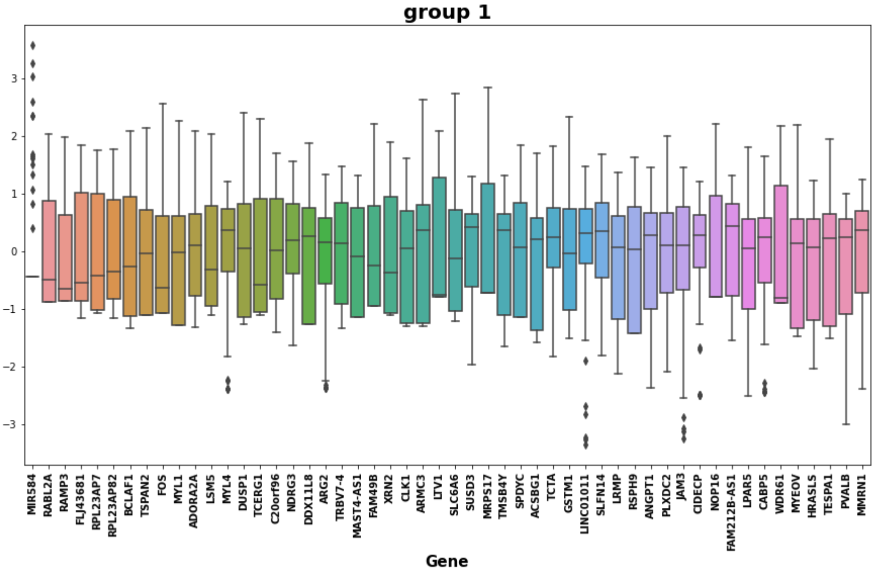

Fig 7: Improved Boxplot of group 1

ROC curve and learning curve

In our study, the ROC (Fawcett, 2006) curves were plotted for each model to evaluate their performance. The model with the highest AUC on the test set will be applied to important gene selection. The ROC curve before improvement is shown below.

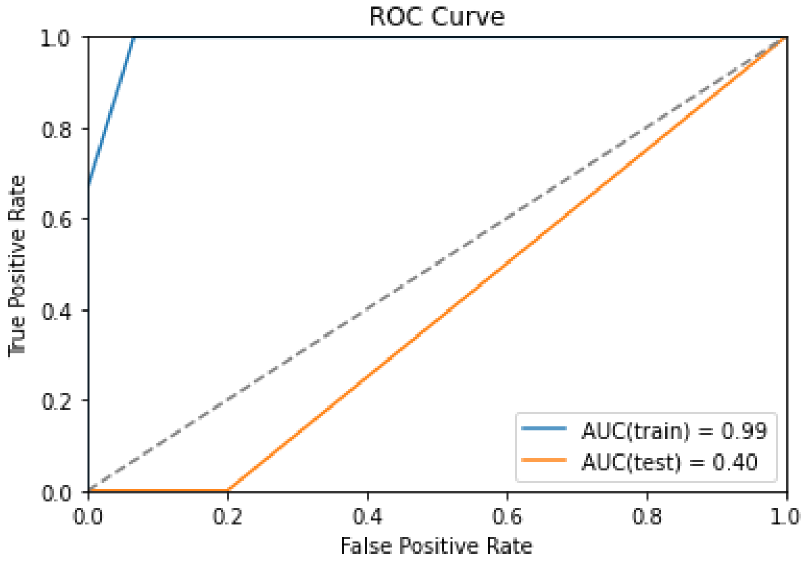

Fig 8: Initial ROC curve

The graph below shows the improved ROC curve. We eliminate the training curve because it is meaningless for evaluation. We perform cross-validation, so we plot curves for each fold with different hues and emphasis the mean result with a larger linewidth.

We add the shadow in the background based on the standard deviation to evaluate the robustness of the model performance. We also calculate the AUCs for each fold and added them to the legend. We also added a grid to make the comparison between each fold easy to carry out. The improvement is based on the Gestalt principles (similarity, continuation, figure/ground).

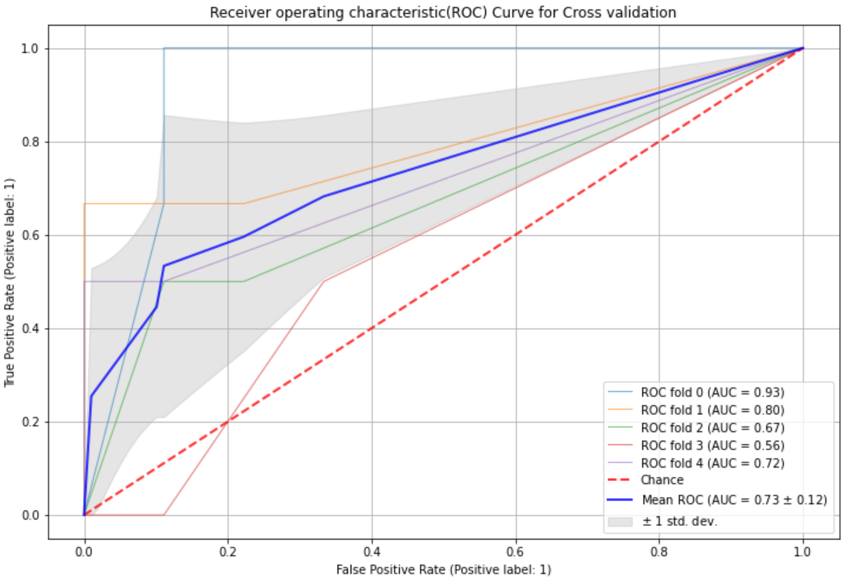

Fig 9: Improved ROC curve

The learning curve is mainly used to evaluate the performance of the model on different sample sizes and their convergence rate in our study. The graph below shows the curve before improvement.

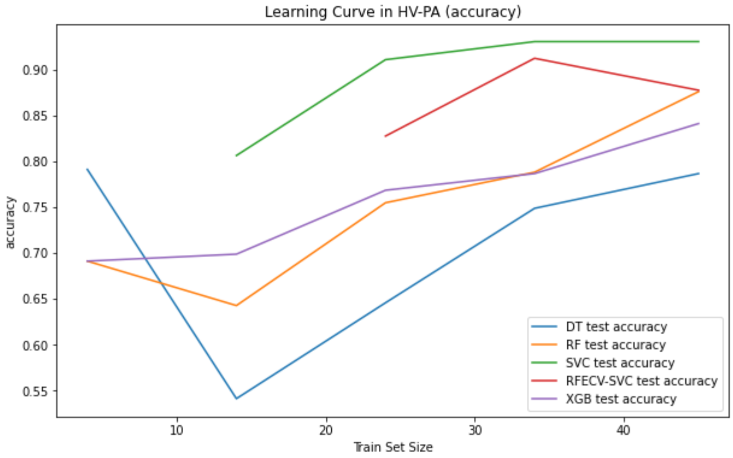

Fig 10: Initial learning curve

The graph below shows the learning curve after improvement. To check the overfitting of the model, we added training curves with the dotted line. The curves with the same hues are for the same model (training and testing). We also added markers and a grid to make comparing different models easy to perform. The improvement is based on the similarity and proximity principles in gestalt principles.

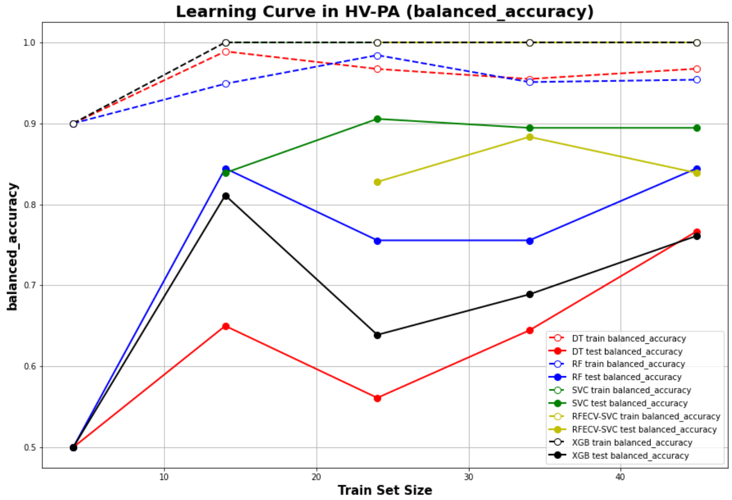

Fig 11: Improved Learning Curve

Dimension reduction and Clustering

To extract the underlying information of DEG, we performed PCA (Daffertshofer, Lamoth, Meijer, & Beek, 2004) for dimension reduction and applied K-means and hierarchical clustering on the DEG. The 2D scatter plot below shows the distribution of samples on the dimension of two PCs and the clusters generated using K-means clustering. No overlap can be observed between clusters, which indicates the two PCs we retain most information of the original data.

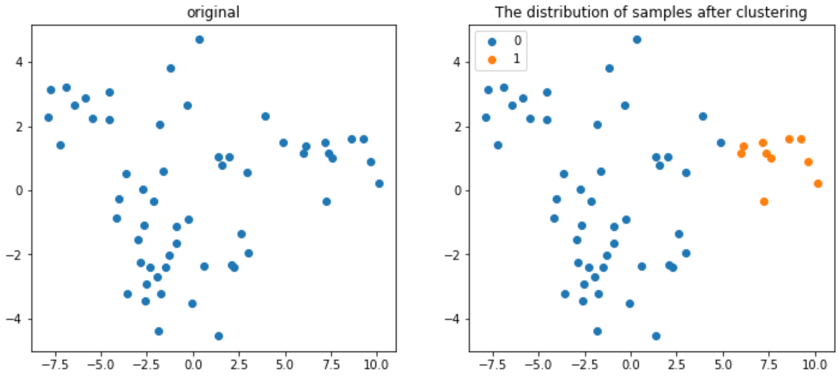

Fig 12: Scatter plot of samples

The dendrogram below is generated using hierarchical clustering. We plot the clusters with the same color as the scatterplot above. From this plot, we can clearly observe the correlation between different genes and the procedure of clustering.

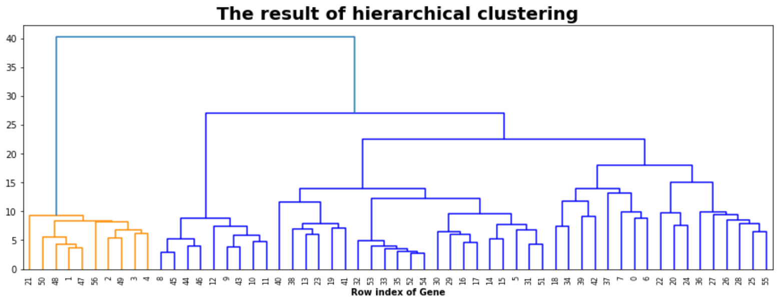

Fig 13: Hierarchical clustering

The 3D scatterplot below is plotted based on the first three PCs and the class label of the data. We can still observe the clusters even in a high-dimension space.

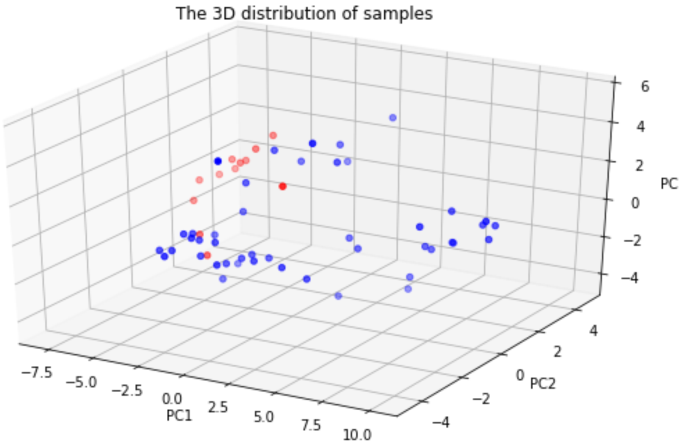

Fig 14: 3D distribution of samples

However, although 3D graphs are helpful for us to search for information in more dimensions, it may confuse the reader. Therefore, we should avoid using 3D graphs in our reports.

Heatmap for chosen variables

We plot the heatmap (Metsalu & Vilo, 2015) for chosen variables to present the clustering relationship of multi-sample or multi-gene expression. The graph below is the heatmap we generate in the first place. Each row represents a gene, and each column represents a sample.

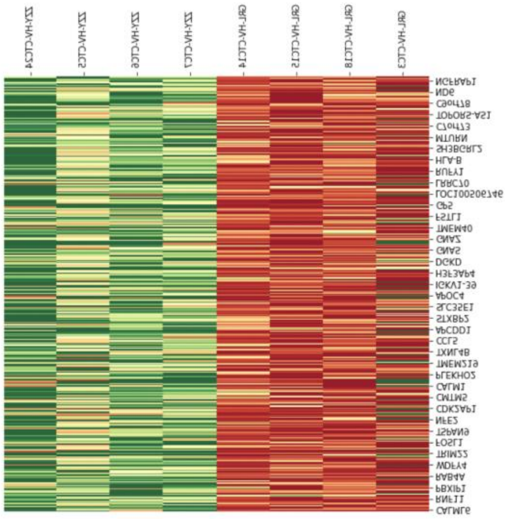

Fig 15: Original Heatmap for chosen variables

The heatmap below is the improved one. We reset the threshold for the DEG selection. To have better visualization of the clusters, we change the colormap and added the color bar. We also added the dendrogram of hierarchical clustering to visualize the correlation between samples and genes. In the improved heatmap, Blue represents relatively low expression and red represents relatively high expression. Those samples or genes with similar expression have been clustered together.

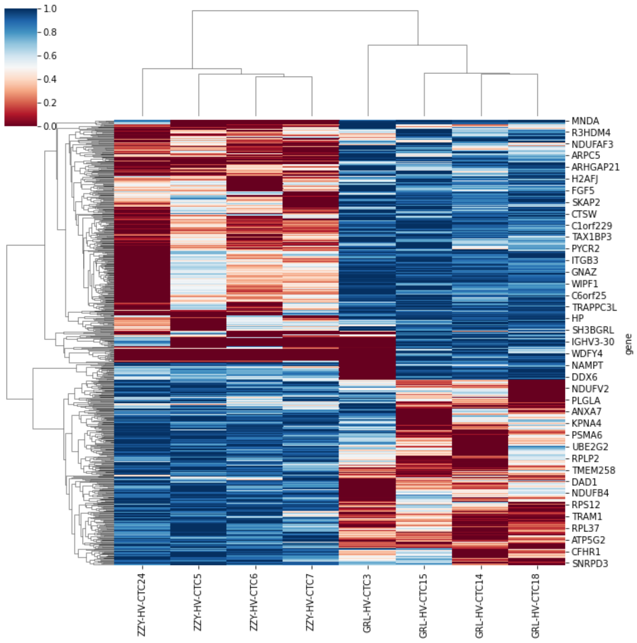

Fig 16: Improved Heatmap for chosen variables

Result Analysis

The normalized data is used for downstream analysis.

As aforementioned, the single cell gene expression profile contains samples with 2 labels: patients and locations. The strong visualization strategy is needed to investigate this informative dataset.

Explore the single gene heterogeneity on location or patient level

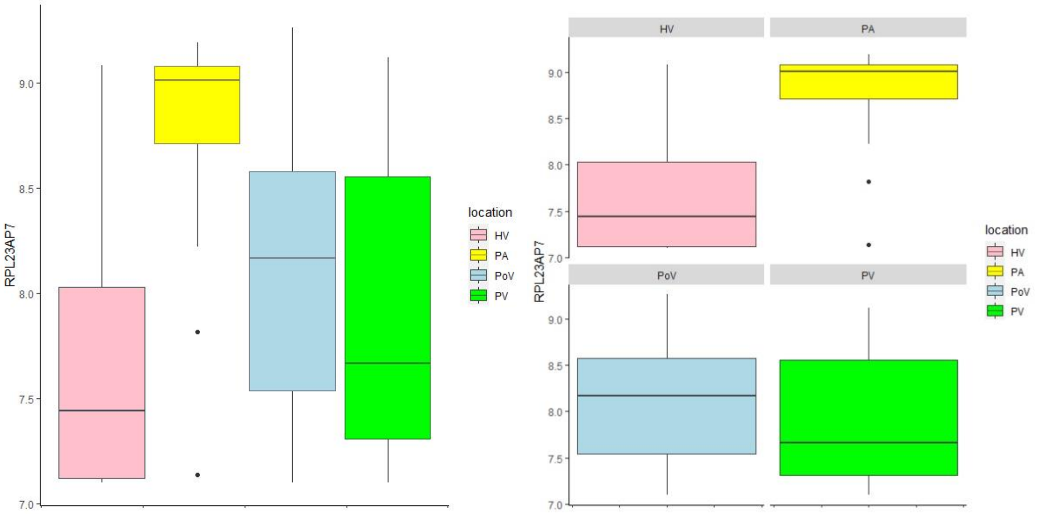

Fig 17: The boxplot on location level with and without facet strategy

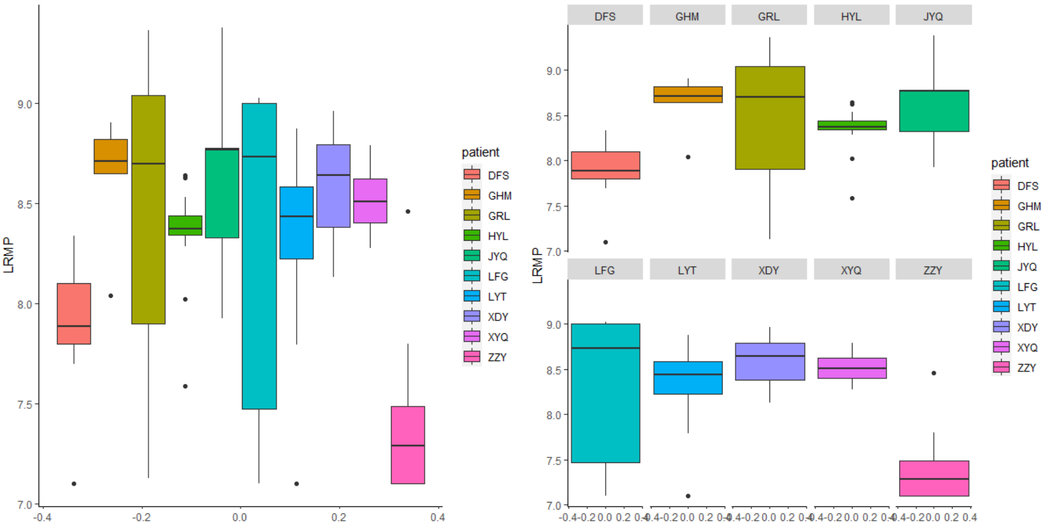

Fig 18: The boxplot on patient level with and without facet strategy

We can utilize facet strategy to organize the figure. And it performs better when the number of classes is large.

Except boxplot, density plot can also achieve the purpose of comparison. Unlike boxplot which highlight the average level of amount, density plot exhibits the mode, or the value that samples gathering.

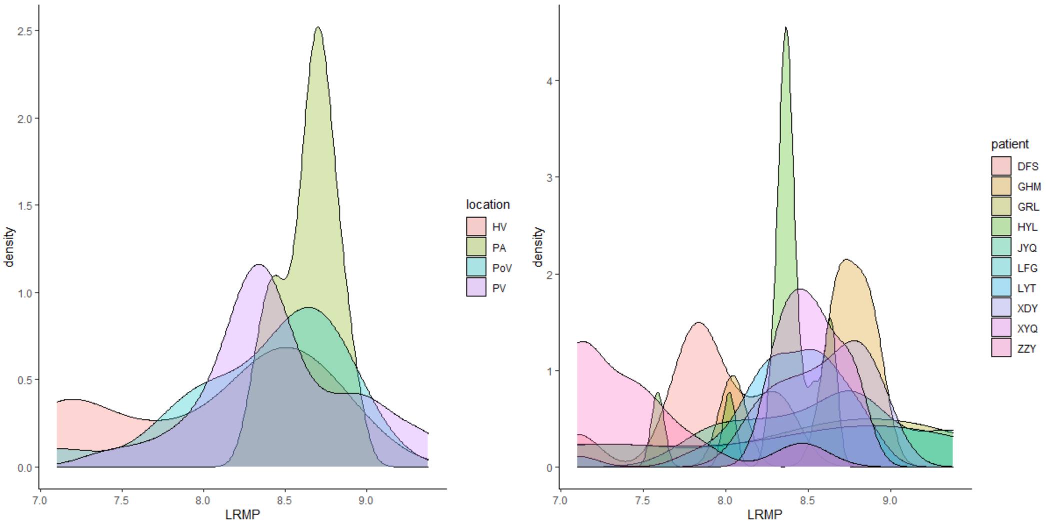

Fig 19: The density plot in location and patient level

The facet strategy is strongly useful when there are more than 2 variables require demonstrated. We can see how a single gene differentially expressed in patient level and location level in only a one figure by applying it. We can designate the which variables mapped to color, depending on whether we aim to emphasize the location level or patient level.

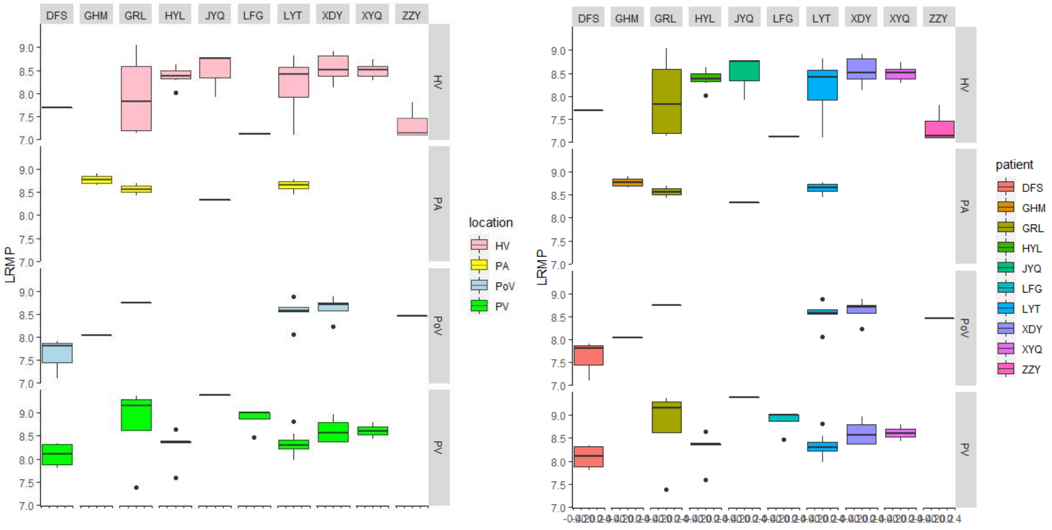

Fig 20: The box plot with facet strategy emphasize location or patient

Explore the correlation of expression level between 2 genes

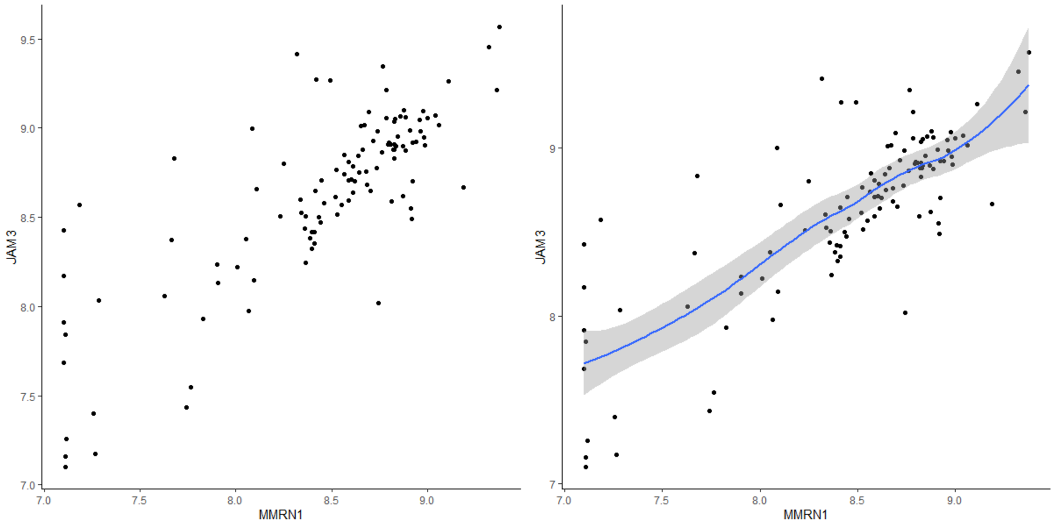

Fig 21: The scatter plot without or with smooth plot

It’s easy to devise a scatter plot to show the correlation (Cohen & Handy, 2001).

But if we use the combination of scatter plot and smooth plot, the correlation between 2 genes will be more perspicuous.

Explore the different expression level between 2 genes among locations and patients

We aim to compare 2 genes among each level simultaneously in one plot. The first pilot is to map continuous variable to color and size.

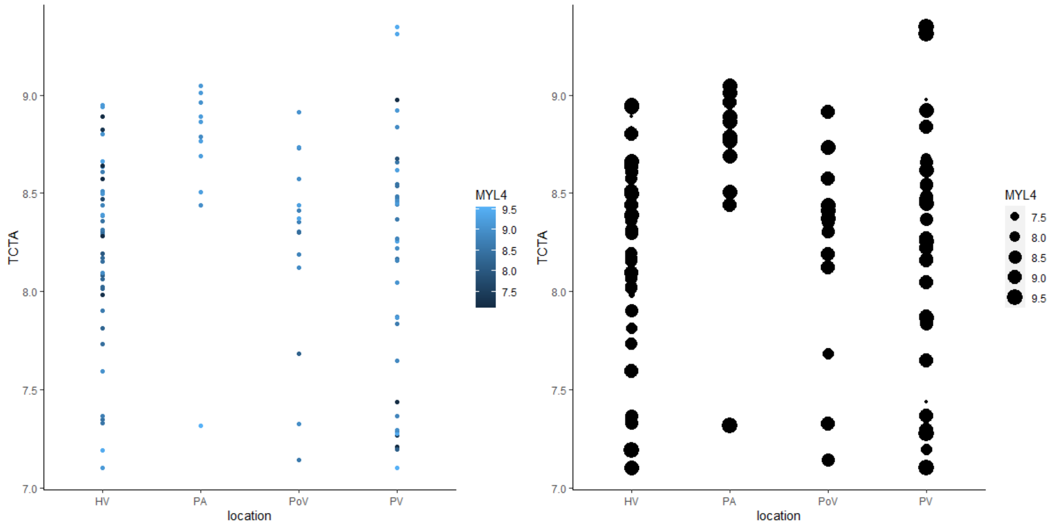

Fig 22: Mapping continuous variable to color and size

The x axis and y axis stand for locations and one of interested gene. The other gene is represented by color(saturation) or size. They are not strong visualization methods for the task.

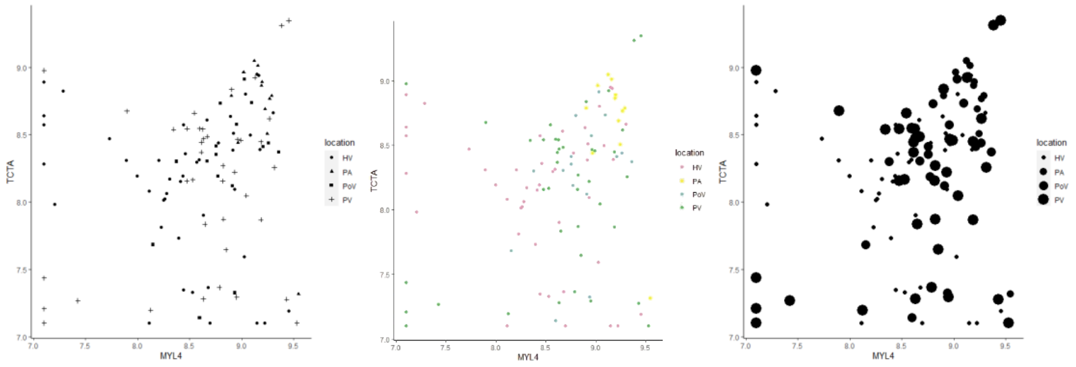

Fig 23: Mapping categorical variable to shape, color and size in a scatter plot

And then, both of genes are projected to x and y axis, and locations are mapped to shape, color and size. Compared with previous figures, they look better. But it still has room to improvement.

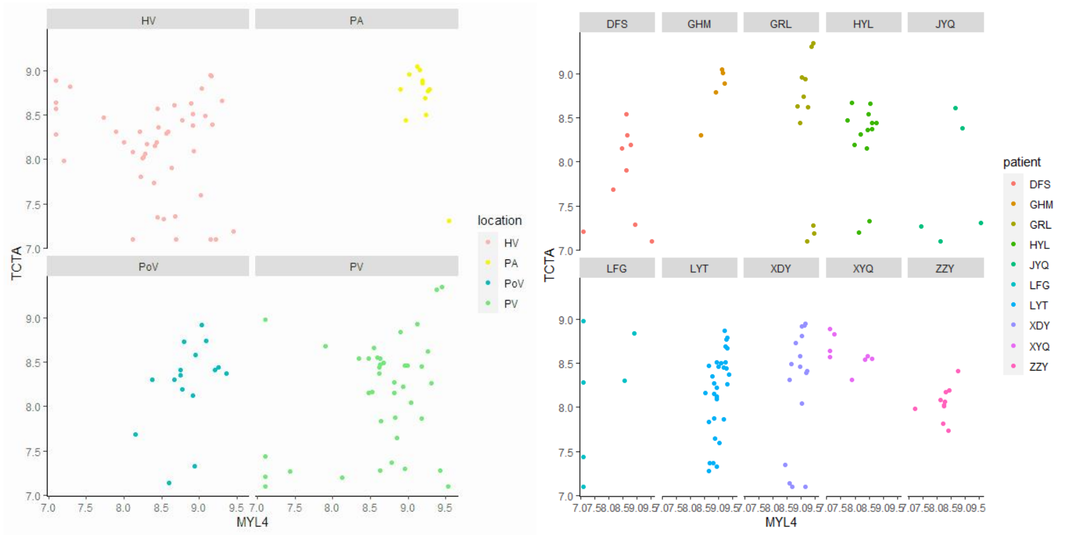

Fig 24: The scatter plot with facet strategy in location or patient level

The combination of facet strategy and color is the optimal solution to show 2 genes expression level among 2 levels.

Explore the sample size of locations and patients

To obtain the insight of which patient or location has abundant data, we investigate the sample size among different level. The count plots are created, which has default setting to use size to express the amount of data.

We've tried to map location to different aesthetics (Meyer & Verrips, 2008). Under this circumstance, as a categorical variable, location is more suitable to be mapped to color aesthetic, although shape can also contribute to discrimination effect to some extent. The underlying reason may be the collaboration of color and size is more efficient.

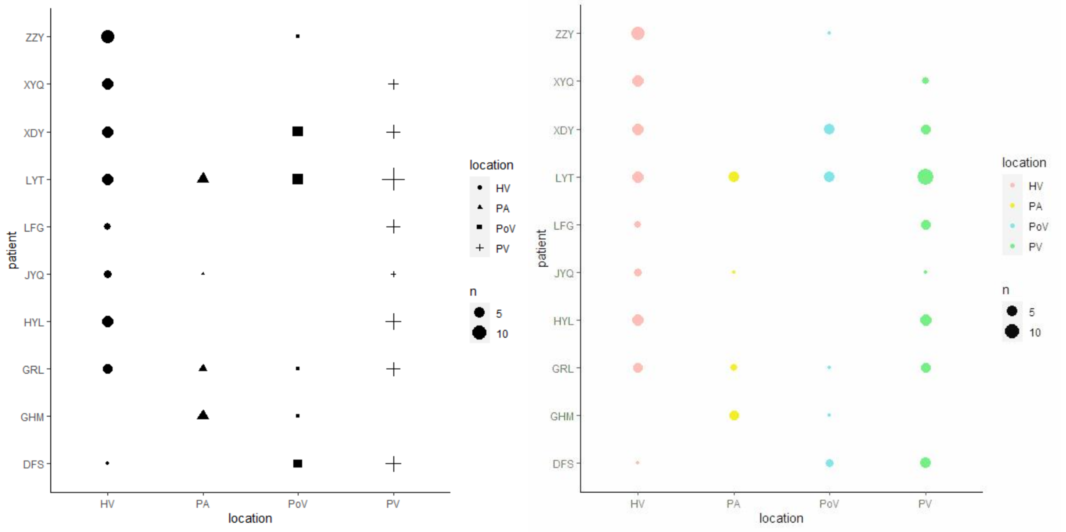

Fig 25: Mapping categorical variable to shape and color in a count plot

The bar chart provides different approach. Rather than points with different size, it shows the number of samples by using height of bar. Therefore, the height of bar gives us perceptual intuition of the sample size derived from each patient or location.

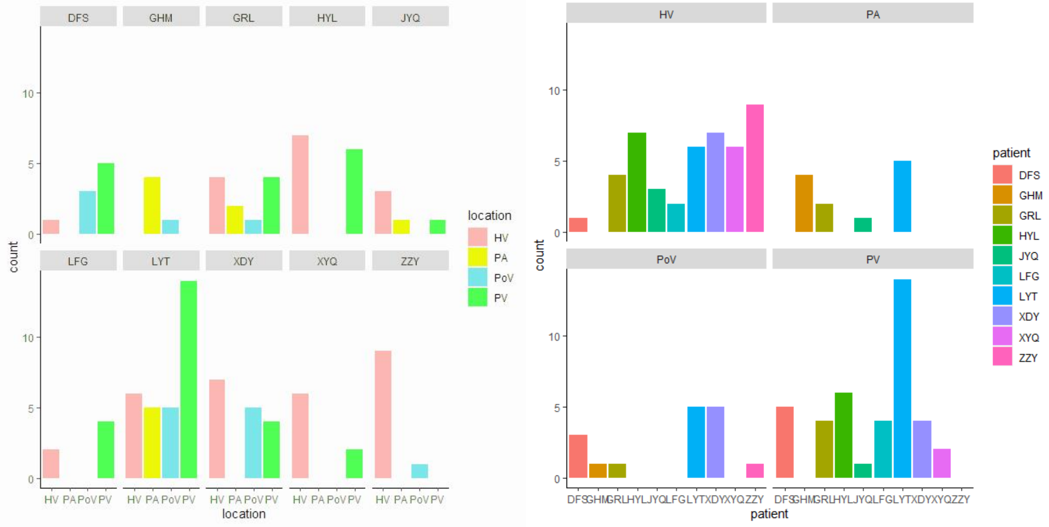

Fig 26: The bar chart with facet strategy on patient or location level

Heatmap and clustering

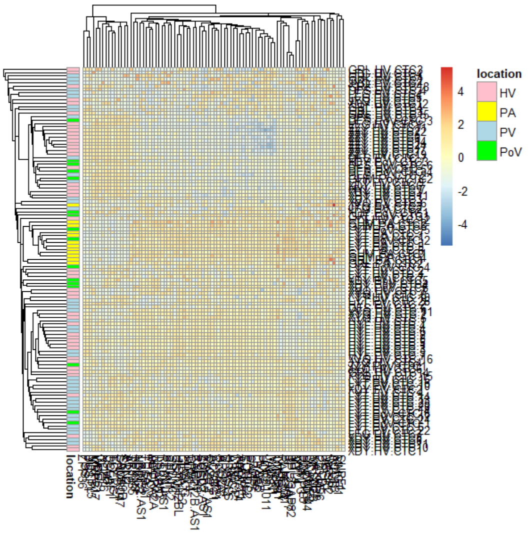

Fig 27: The original clustered heatmap

The initial version of heatmap is constructed with entire DEG. the issue if the map is apparent: ①The column name and row name are messy and overlapped. ②The hue of expression level contains 3 colors and the resultant plot mainly filled with yellow, which confound the reader to some extent. ③The number of genes may be large to the plot and some redundant gene can be removed.

According to the issues, we improve the heatmap by: ①Select only the important DEG recognized by Machine Learning for the heatmap. ②Conceal the row names which represent the name of samples, and rotate the columns name with 45 degrees. ③Adjust the hue of expression level.

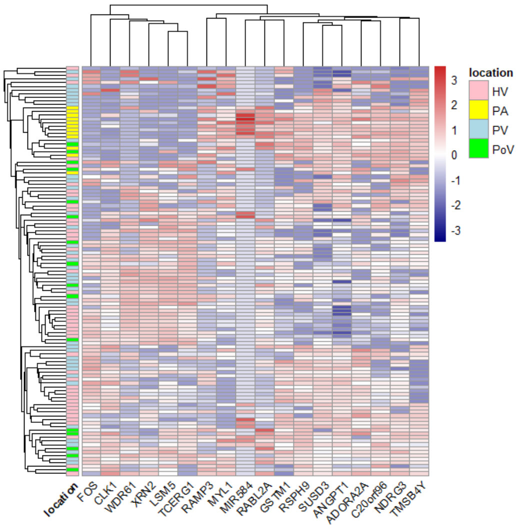

Fig 28: The improved clustered heatmap

The latest version of heatmap can demonstrate the important information strongly. We can see not only the clustering tendency in patient level, but also the correlation among gene expression level.

Summary

Our goal has always been to communicate clearly and accurately. Therefore, this paper shows a visualization of the work of different steps in the whole process of building our model.

The first thing we need to do is unify the colors. Colors are beautiful and confusing, so unifying colors is also the basis for delivering the right message. All the colors representing the four regions in this paper are the same set of colors, and the 10 patients are also represented by another set of ten colors. By controlling the color, we can determine the purpose of the picture at first glance, such as whether it wants to compare patients, or compare positions.

In addition to this, we also need to ensure that the labels are clear and accurate, which is also a necessary step when improving the image. We need to ensure that our data messages are understood by readers and that there are no difficulties in reading them.

Different images serve different purposes, for example Venn (Gao, Yu, & Cai, 2021) diagrams are good at showing the intersection and union of data. It is very helpful for understanding the similarities and differences between different groups of data, but it cannot show the distribution of data within the group. Therefore, in different steps, we need to draw different images according to the requirements, so as to achieve our purpose. In this article, we also use a lot of data to express images, such as scatter plots, boxplots, violin plots, cluster plots, and heatmaps. We use them with care, making their use reasonable and accurate.

Reference

Cohen, A. J., & Handy, N. C. (2001). Dynamic correlation. *Molecular Physics, 99*(7), 607-615.

Daffertshofer, A., Lamoth, C. J., Meijer, O. G., & Beek, P. J. (2004). PCA in studying coordination and variability: a tutorial. *Clinical biomechanics, 19*(4), 415-428.

El-Serag, H. B., & Mittal, S. (2013). *Epidemiology of HCC: consider the population.* Journal of clinical gastroenterology.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters, 27*(8), 861-874.

Gao, C. H., Yu, G., & Cai, P. (2021). GgVennDiagram: an intuitive, easy-to-use, and highly customizable R package to generate venn diagram. *Frontiers in Genetics*, 1598.

Metsalu, T., & Vilo, J. (2015). ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. *Nucleic acids research, 43*(W1), W566-W570.

Meyer, B., & Verrips, J. (2008). Aesthetics. *In Key words in religion, media and culture* , 36-46.

Upstill-Goddard, R., Eccles, D., Fliege, J., & Collins, A. (2013). Machine learning approaches for the discovery of gene–gene interactions in disease data. *Briefings in bioinformatics, 14*((2)), 251-260.

Zhang, R., Ou, H. Y., & Zhang, C. T. (2004). DEG: a database of essential genes. *Nucleic acids research*, D271-D272.

Contribution

Name	Contribution
Ye Shuchen	Draw the picture of Data Display and DEG Display Write the part of the Data Display and DEG Display Write the introduction and summary Integrate articles
Zhang Zhu	Draw and write: Explore the single gene heterogeneity on location or patient level Explore the correlation of expression level between 2 genes Explore the different expression level between 2 genes among locations and patients Explore the sample size of locations and patients Heatmap and clustering
Han Wenhao	Improve the boxplot for feature distribution Improve the visualization of ROC curve and learning curve Perform dimension reduction and clustering result visualization Improve the heatmap for the clustering of DEG in location 1 Write the report
Li Guanchen	Plot Donut chart to see percentage of genes in each location Plot violin plot to check density and distribution of gene expression data after normalization in each location Plot Venn diagram to check intersections of DEGs