Consulting Project

Faculty Supervisor: Masanao Yajima
Teaching Fellow: Shiwen Yang
Group Member: Huaijin Xin, Chenghao Xia, Bolong Xian

2023-12-19

Introduction

The topic of this consulting project is gene distribution in Entorhinal Cortex. Entorhinal Cortex is anatomically positioned between the neocortex and the hippocampus, and its major role is to bridge information exchange between the two regions. Our client, Ana Morello who is a graduate student at department of Anatomy & Neurobiology in School of Medicine in Boston University, recently is doing research about the gene distribution of cells in EC region. They took a monkey's brain and cut it into slices to have several layers of EC section.

She utilized a technique called in-situ hybridization to dye different genes into different colors. In-situ hybridization is a powerful technique used in molecular biology to detect and localize specific DNA or RNA sequences within a tissue section or cell sample. This method involves hybridizing a labeled complementary DNA or RNA probe to the target nucleic acid sequence within the tissue or cells. The probe's label, which can be radioactive or fluorescent, allows for the visualization of the hybridization location, thereby indicating where the specific sequences of interest are expressed within the sample. The specific technique she utilizes is called the RNAscope Multiplex Fluorescent Assay v2 which is a more advanced version of in-situ hybridization designed specifically for the simultaneous detection of multiple RNA targets within a single sample. This technique employs fluorescent labeling, enabling researchers to visualize and quantify the expression of several different RNA molecules at once. The "multiplex" nature of the assay allows for the co-localization of different RNA species within the same sample, providing a comprehensive understanding of gene expression patterns and interactions. Different fluorescent dyes for multiplex fluorescence imaging: Opal 520, 570, 620, 690. Number represents the wavelength in nanometer of light and those also represent different genes in the datasets. The measurement she got is fluorescent intensity which is A measure of the amount of fluorescence emitted by a sample. Fluorescence is a phenomenon where certain molecules absorb light (photons) at one wavelength and then re-emit light at a longer wavelength. Higher the Fluorescent Intensity means higher the concentration of certain gene in the selected cell.

The datasets we get are 3 layers of different fluorescent intensity measures from the reflection of different wavelengths (520, 570, 620, 690) in different cells and the datasets also consists of the horizontal distance between the cell and the edge of the slice of the EC region. And it also has a column which represents which of the 4 genes is positive for this cell. There are still lots of variables in the raw data that we did not use in this project such as the x axis and y axis of the cell.

The goal of the project is firstly count the number of positive cells for different genes, secondly show the correlation between different genes, thirdly show the distribution of four type of genes, and lastly find the relationship of cells between each layers.

Data Cleaning

We have divided the three layers into two datasets: one comprises cells containing Opal_520, while the other includes all cells, whether or not they contain Opal_520. Typically, we utilize the dataset where all cells contain Opal_520. Here is an example showing the first five rows of this dataset:

Class	Opal_520	Opal_570	Opal_620	Opal_690	Distance
520:570:690	0.3483	0.1596	0.0225	0.1164	2871.8301
520:570:690	0.2152	0.1041	0.0196	0.1136	2866.8936
520:570:690	0.5518	0.0258	0.016	0.2296	2861.261
520:690	0.4816	0.0202	0.02	0.3153	2868.6372
520:570	0.2459	0.1088	0.0211	0.0229	2918.8682

In a separate dataset, we assess the presence of genes in each cell, incorporating four additional columns with boolean outputs. Below is an illustration featuring the first five rows of this particular dataset:

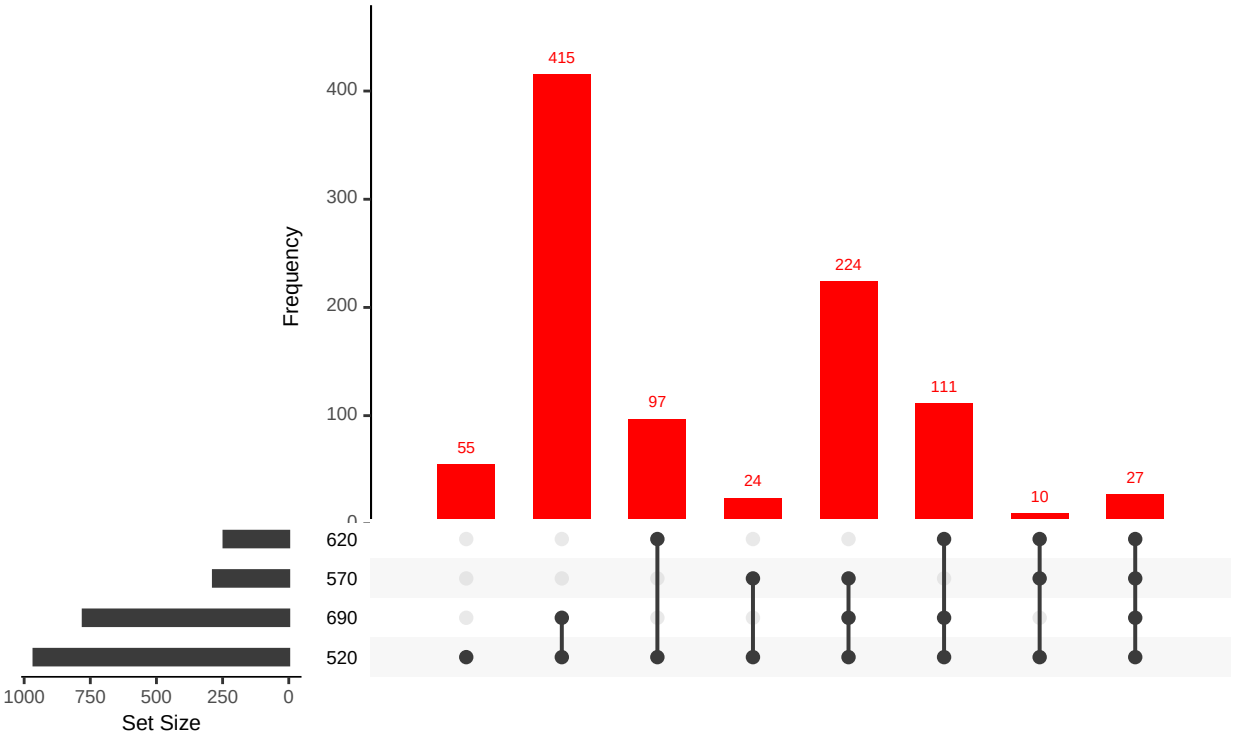
MFI520	MFI570	MFI620	MFI690	dist	IND520	IND570	IND620	IND690
0.3483	0.1596	0.0225	0.1164	2871.8301	TRUE	TRUE	FALSE	TRUE
0.2152	0.1041	0.0196	0.1136	2866.8936	TRUE	TRUE	FALSE	TRUE
0.5518	0.0258	0.016	0.2296	2861.261	TRUE	TRUE	FALSE	TRUE
0.4816	0.0202	0.02	0.3153	2868.6372	TRUE	FALSE	FALSE	TRUE
0.2459	0.1088	0.0211	0.0229	2918.8682	TRUE	TRUE	FALSE	FALSE

Both datasets undergo a cleaning process wherein values associated with fluorescent intensity equal to 0 are eliminated. Additionally, we have renamed certain column names for better clarity and understanding. When examining the fluorescent intensity for each gene, we employ the logarithm to enhance our ability to visualize the distribution.

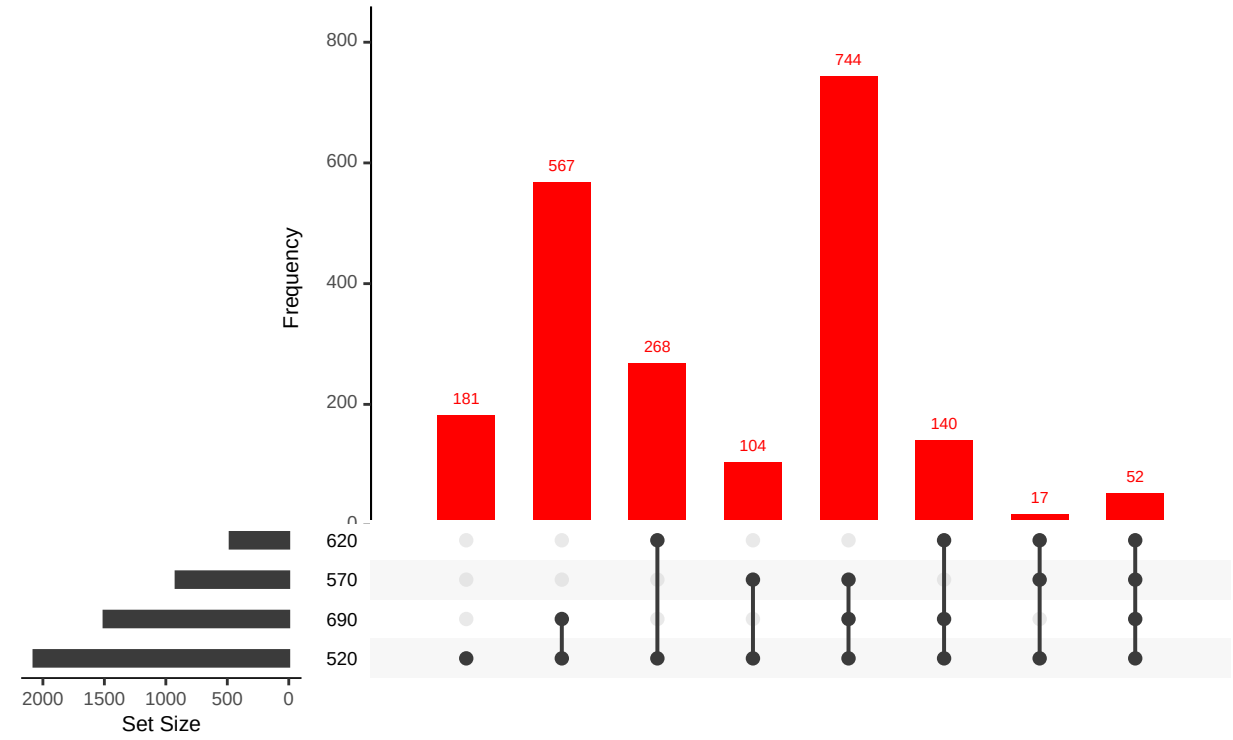
Visualization

Upset Plot

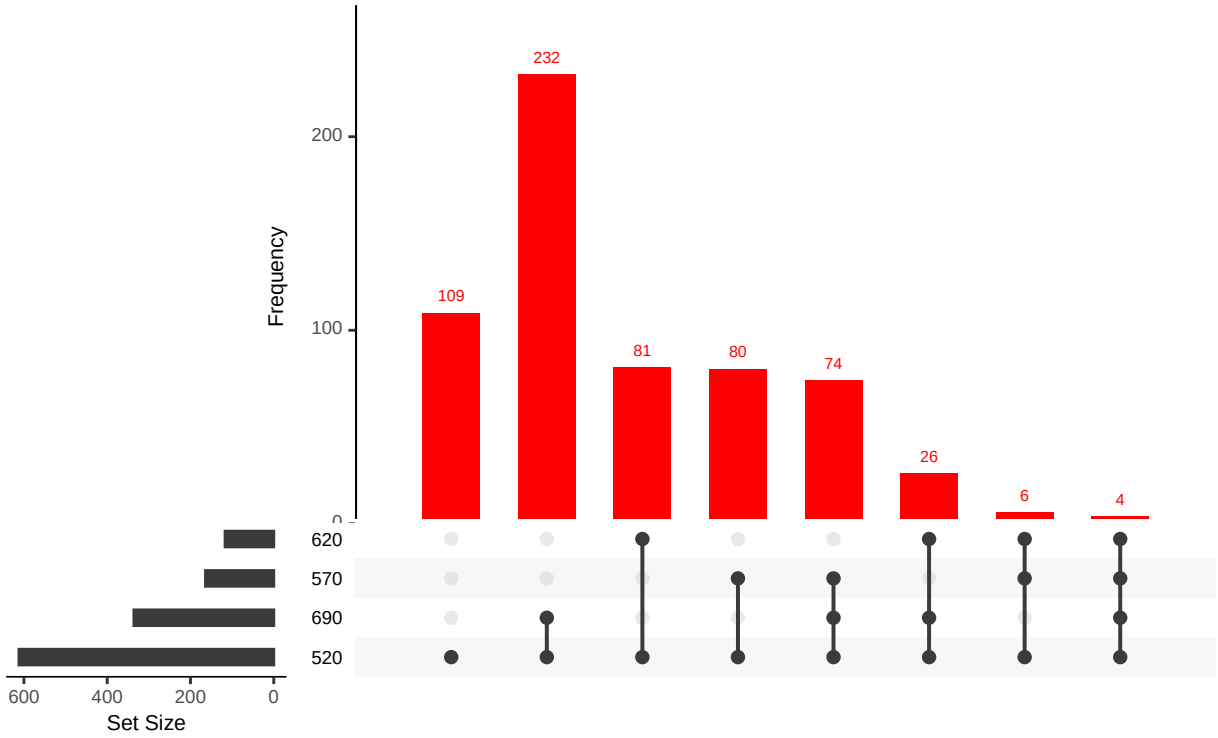
We employ the UpSet plot, a powerful visualization tool, to gain insights into the distribution of elements across three distinct layers in our dataset. The UpSet plot provides a comprehensive overview of intersecting sets, showcasing the frequency and relationships among them. Specifically, it allows us to explore how various combinations of elements from the three layers contribute to the overall distribution, enabling a nuanced understanding of patterns and overlaps in the data.



The above figure illustrates the distribution for layer12_4. By employing data that includes only cells containing Opal_520, the count of Opal_520 serves as a direct indicator of the number of cells in the dataset, totaling around 1000. Notably, within this layer, the cell combination 520:690 exhibits the highest frequency, with a count of 415.



The figure above depicts the distribution for layer12_5. Utilizing data that exclusively includes cells containing Opal_520, the count of Opal_520 directly represents the number of cells in the dataset, totaling approximately 2000. Notably, within this layer, the cell combination 520:570:690 exhibits the highest frequency, with a count of 744.

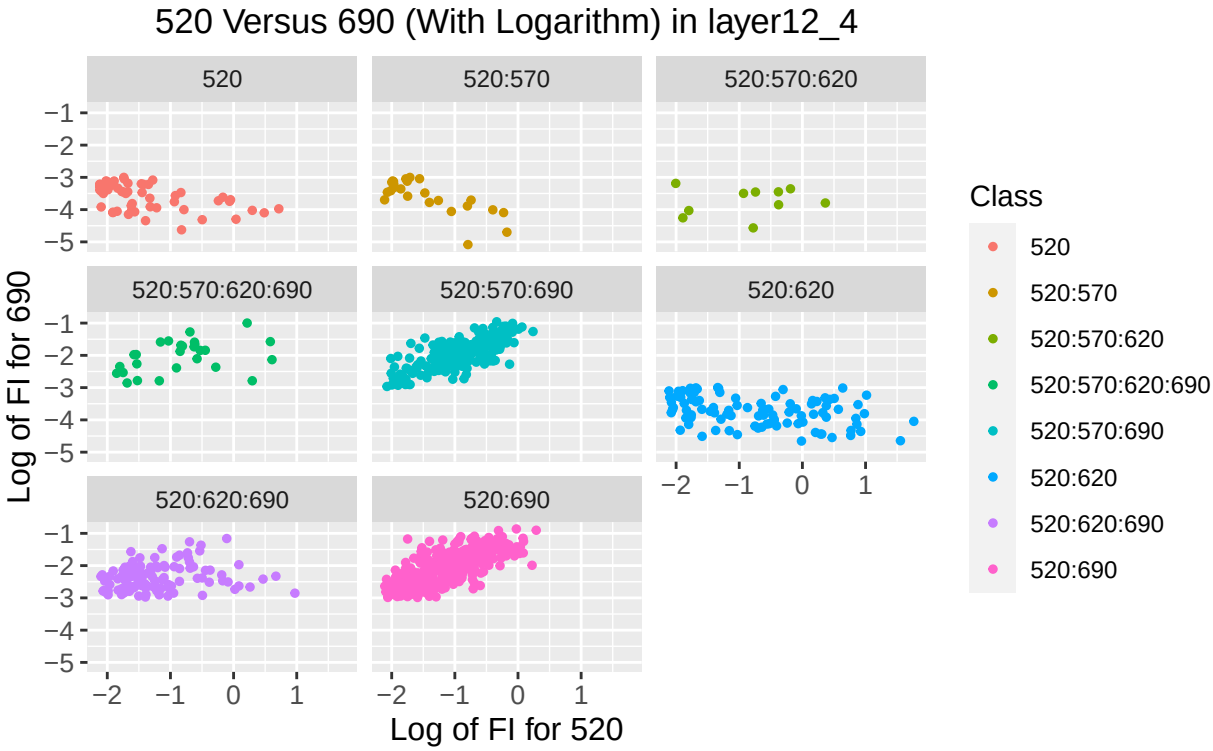


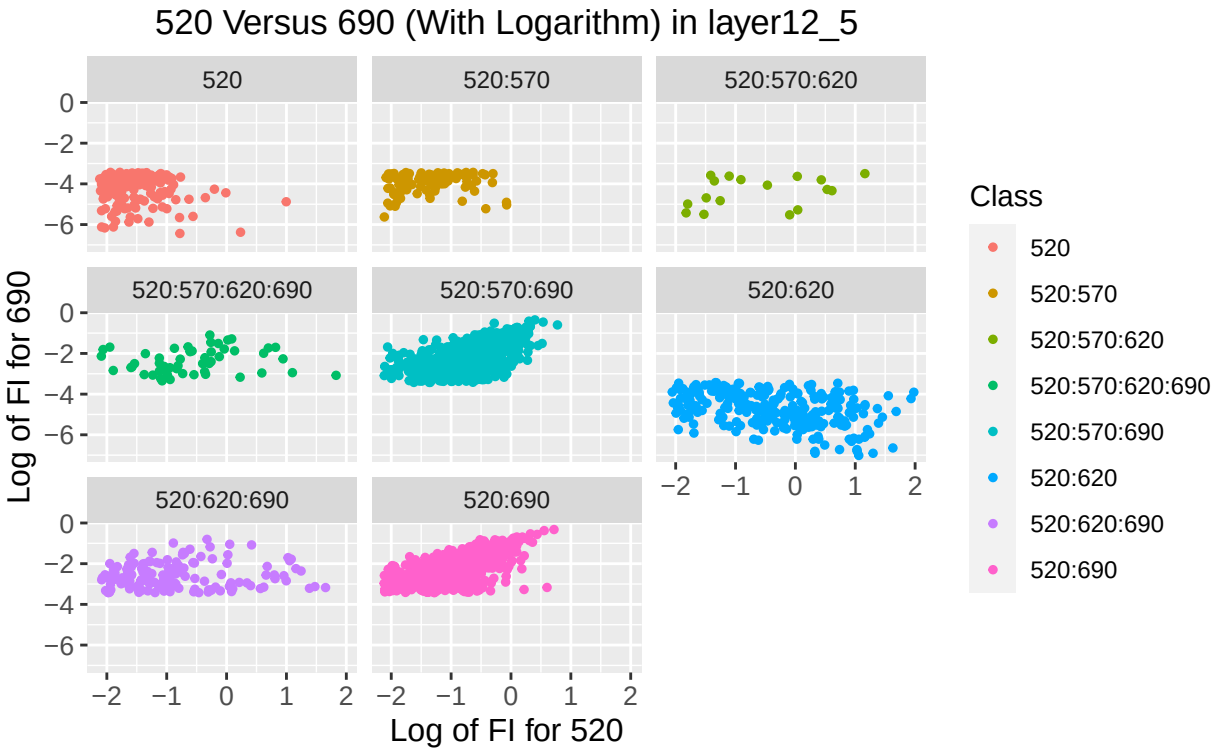
The figure above illustrates the distribution for layer20_5. Utilizing data comprising cells with Opal_520, the count of Opal_520 directly corresponds to the number of cells in the dataset, totaling approximately 600. Notably, the cell combination 520:690 exhibits the highest frequency, with a count of 252 within the layer.

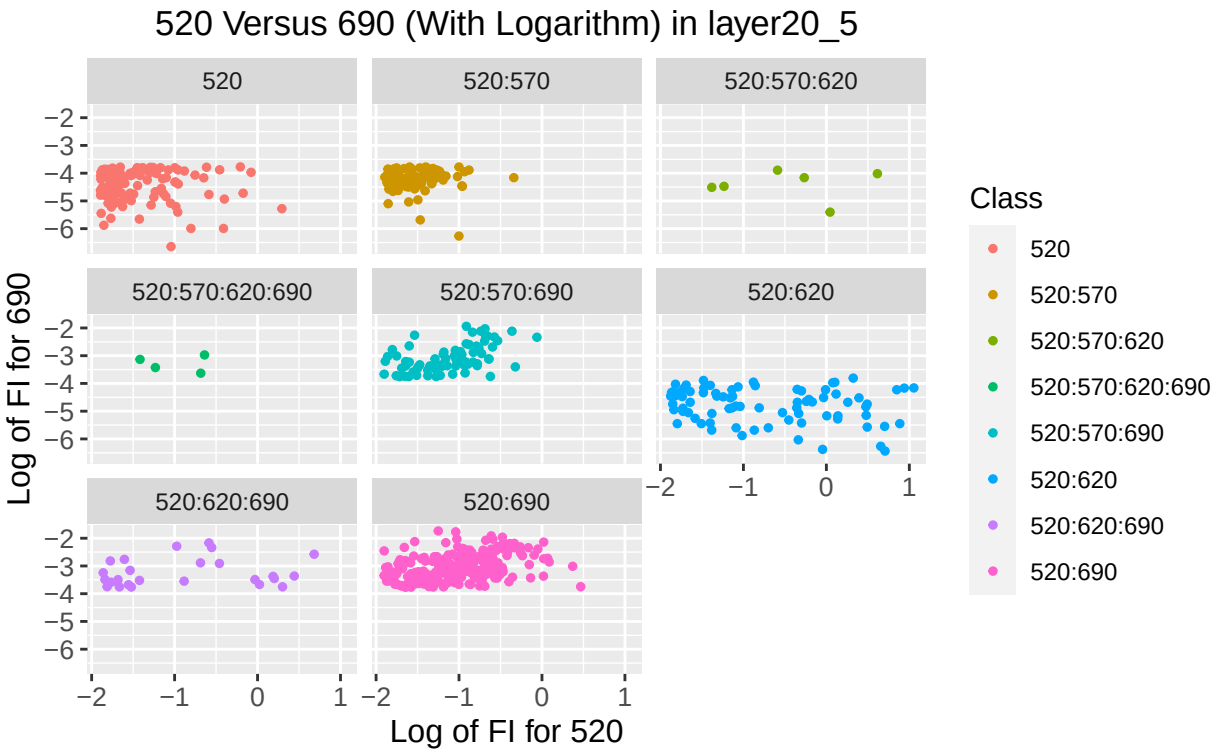
In summarizing the insights from these three figures, a notable observation is that Opal_690 appears more frequently than Opal_570 and Opal_620 across layer20_5, layer12_4, and layer12_5. Both layer20_5 and layer12_4 exhibit 520:690 with the highest frequency in their respective layers. However, in layer12_5, while 520:690 boasts a substantial frequency, 520:570:690 holds the highest frequency value. Interestingly, Opal_520 and Opal_690 frequently co-occur. Moving forward, our analysis aims to delve deeper into understanding the relationship between Opal_690 and Opal_520.

Correlation

We employ a scatterplot to visually analyze the distribution of Opal_520 and Opal_690. In this plot, the x-axis represents the logarithmic values of the fluorescent intensity for Opal_520, while the y-axis depicts the logarithmic values of the fluorescent intensity for Opal_690. This visualization allows us to explore the relationship and potential patterns between the two variables in a quantitative manner.


520 Versus 690 (With Logarithm) in layer12_4
520
520:570
520:570:620
520:570:620:690
520:570:690
520:620
520:620:690
520:690
Log of FI for 690
Log of FI for 520
Class
520
520:570
520:570:620
520:570:620:690
520:570:690
520:620
520:620:690
520:690


520 Versus 690 (With Logarithm) in layer12_5
520
520:570
520:570:620
520:570:620:690
520:570:690
520:620
520:620:690
520:690
Log of FI for 690
Log of FI for 520
Class
520
520:570
520:570:620
520:570:620:690
520:570:690
520:620
520:620:690
520:690



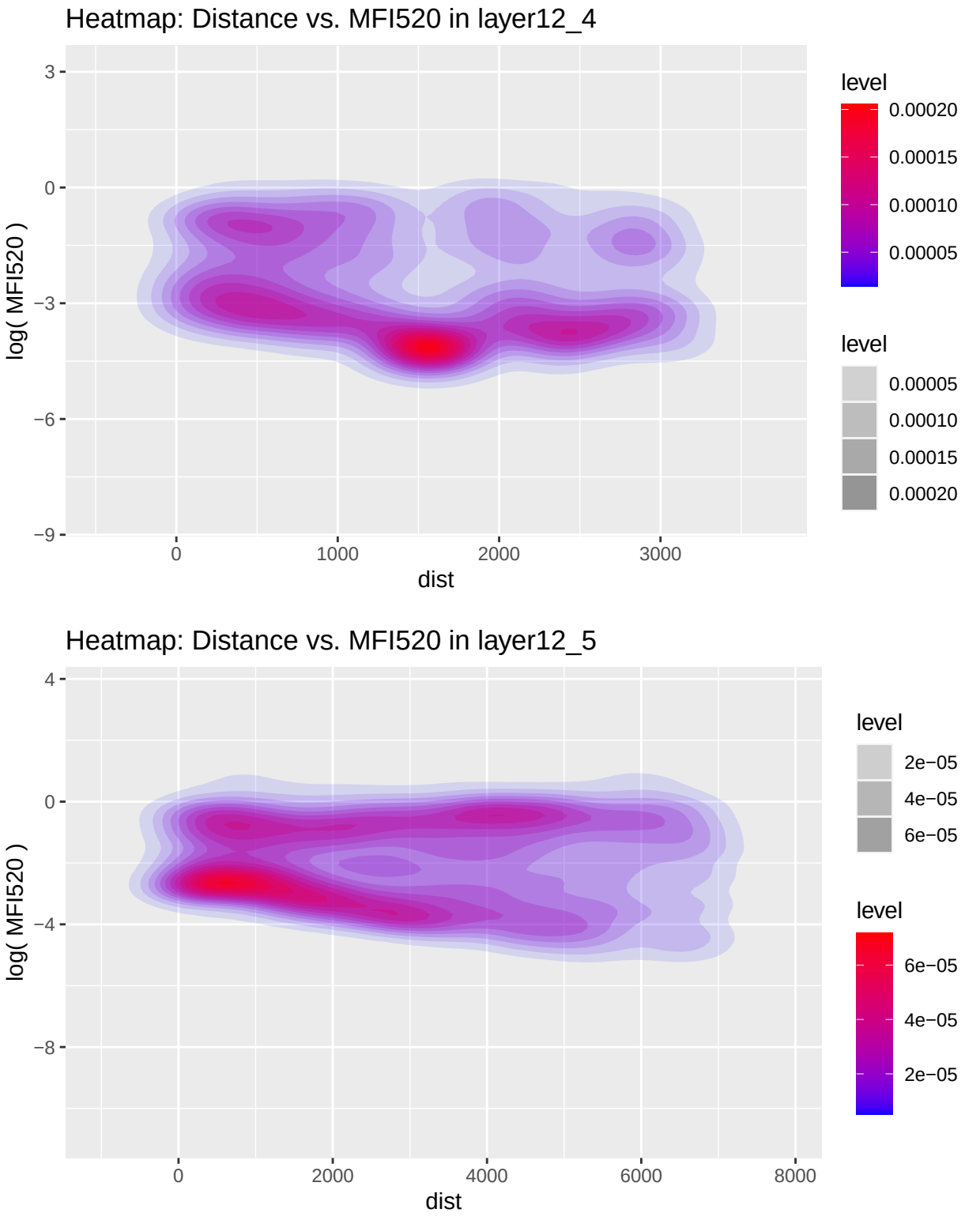
Based on the figures above, it is evident that cells containing Opal_690 tend to exhibit higher values for Opal_690, which aligns with expectations. In both layer12_4 and layer12_5, the cells with combinations 520:570:690 and 520:690 demonstrate a positive correlation between Opal_520 and Opal_690. As the value of Opal_520 increases, there is a corresponding increase in the value of Opal_690. However, in layer20_5, a notable correlation between Opal_520 and Opal_690 is not readily apparent. The presence of a subtle positive correlation in layer20_5 raises the question of whether this correlation is a common feature across all layers or if it is specific to layer12_4 and layer12_5. Further investigation is required to ascertain whether the observed correlation is consistent across all layers or if it exhibits variation, with layer12_4 and layer12_5 demonstrating a more pronounced association compared to layer20_5.

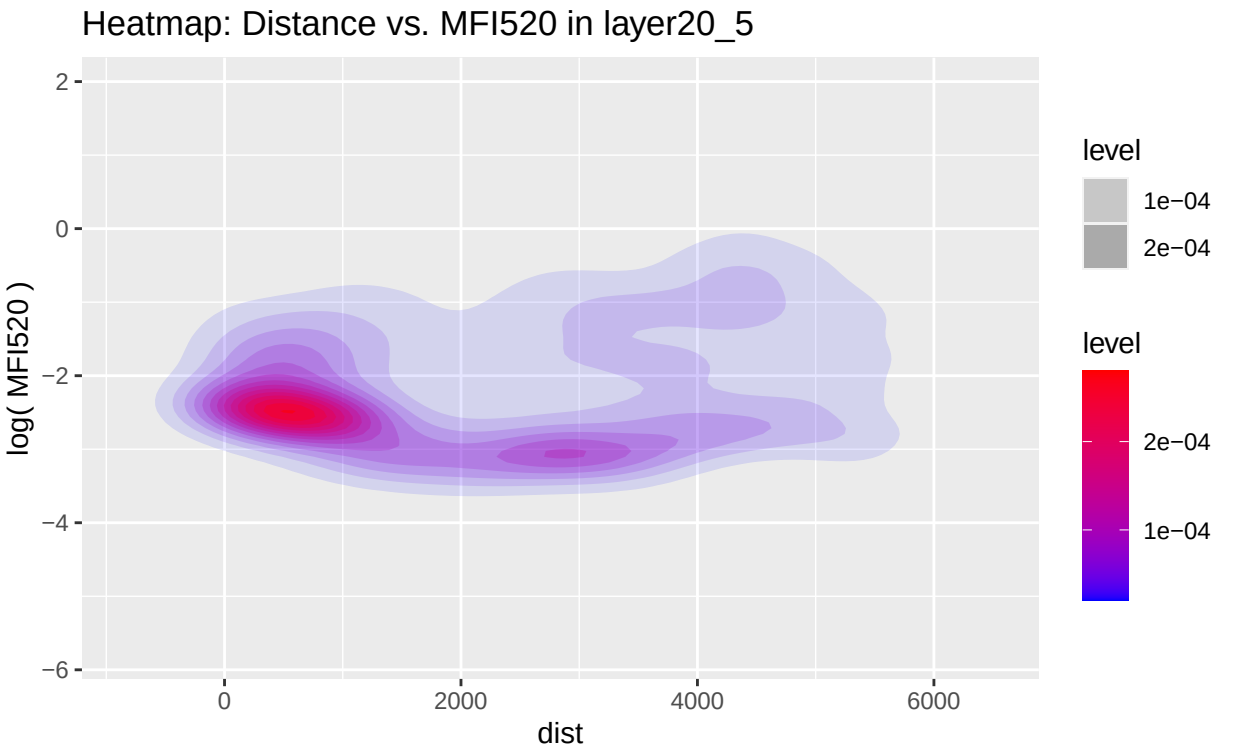
Heatmap

We will construct heatmaps to illustrate the correlation between genes and distance.

Heatmap for distance and one gene

These heatmaps depict the correlation between distance and the log-transformed mean fluorescent intensity (MFI) of a specific gene in a given dataset. The selection of Opal_520 is motivated by its strong correlation with Opal_690. Opting for Opal_520 is advantageous as it serves as the indicator gene for the region of interest. In the visual representation, the x-axis corresponds to the Distance values, and the y-axis portrays the logarithmic values of the fluorescent intensity for Opal_520.

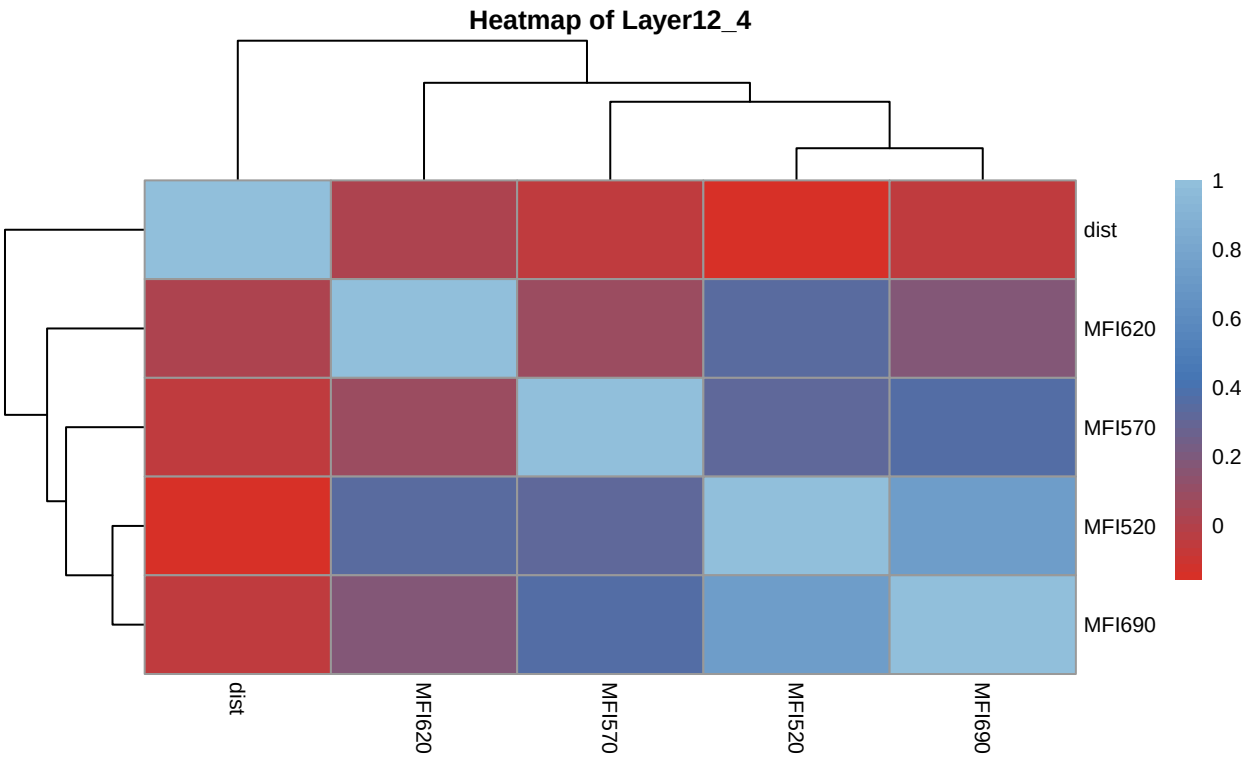

Heatmap: Distance vs. MFI520 in layer12_4
log(MFI520)
dist
3
0
−3
−6
−9
0
1000
2000
3000
level
0.00020
0.00015
0.00010
0.00005
level
0.00005
0.00010
0.00015
0.00020
Heatmap: Distance vs. MFI520 in layer12_5
log(MFI520)
dist
4
0
−4
−8
0
2000
4000
6000
8000
level
2e−05
4e−05
6e−05
level
6e−05
4e−05
2e−05

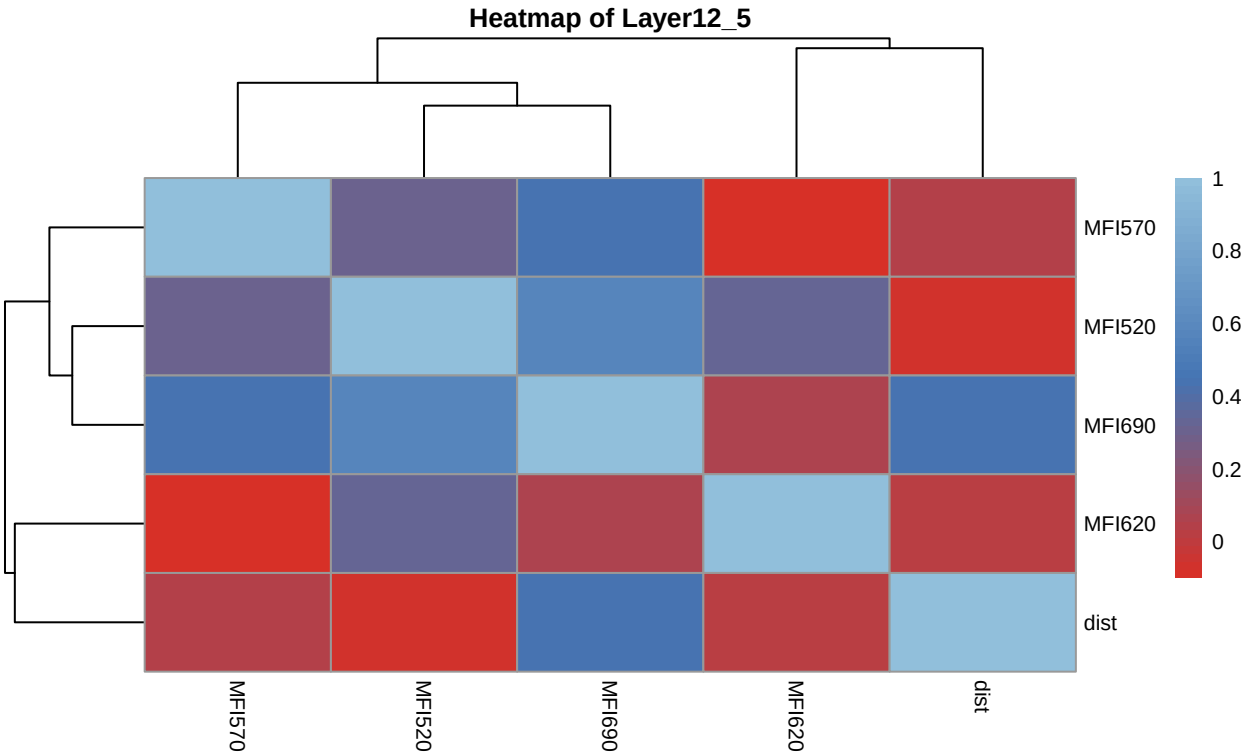


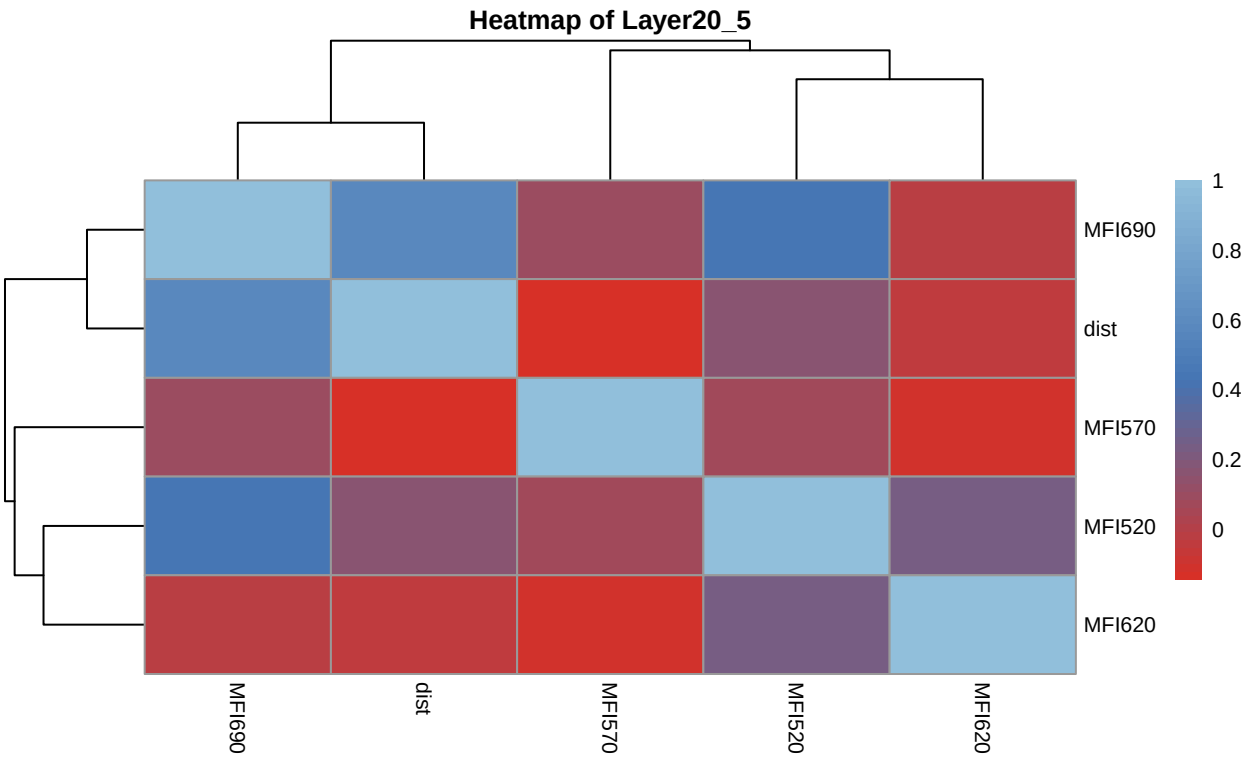
The distribution of Opal_520 appears to vary across different layers. In layer12_4, the plot exhibits a high density centered at coordinates (x=1500, y=-4.5) with a subtle vertical symmetry. Contrarily, for layer12_5 and layer20_5, the highest density is skewed towards the left, particularly around x=1000. Notably, there are distinctions between layer12_5 and layer20_5. In layer12_5, there is a slight horizontal symmetry, whereas layer20_5 lacks a discernible trend. To generate additional insights using the provided R code, users can experiment with different datasets, gene numbers, and layer names by adjusting the function parameters such as "**dataset**", "**gene_number**", and "**layer_name**" in the dist_intensity_heatmap function.

Heatmap for all variables

We also came up with heatmaps to indicate the relationship of different genes. And we can see that there is some difference between the correlations among 3 layers.


Heatmap of Layer12_4
dist
MFI620
MFI570
MFI520
MFI690
1
0.8
0.6
0.4
0.2
0

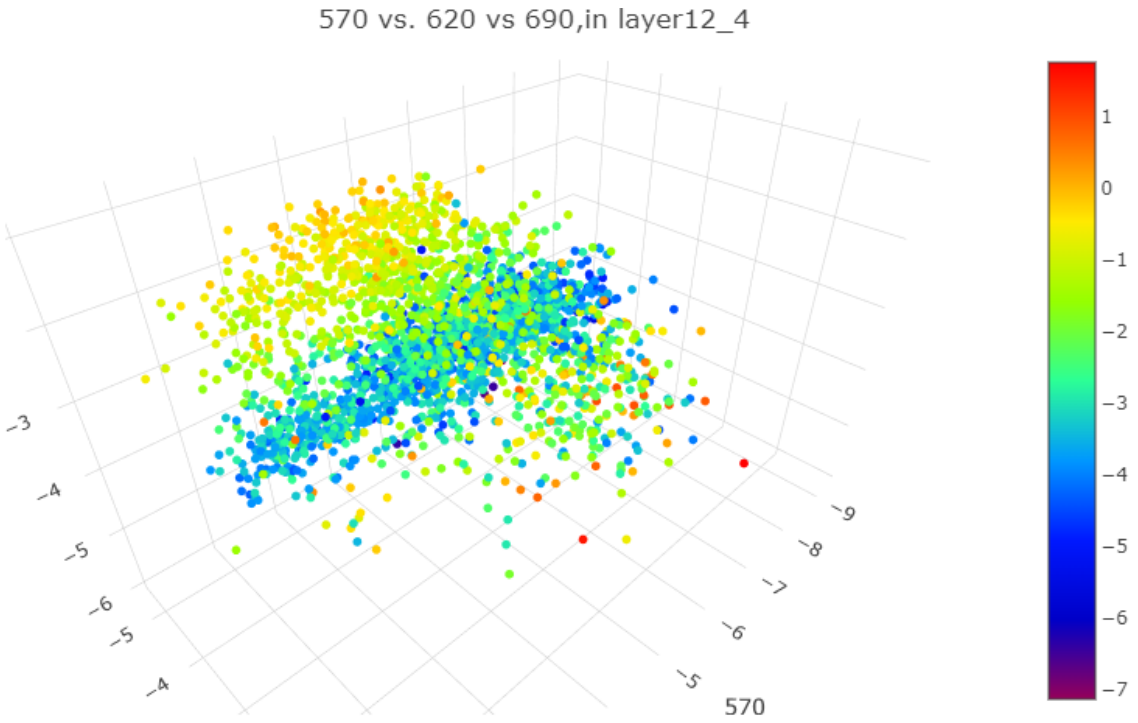

Heatmap of Layer12_5
MFI570
MFI520
MFI690
MFI620
dist
1
0.8
0.6
0.4
0.2
0

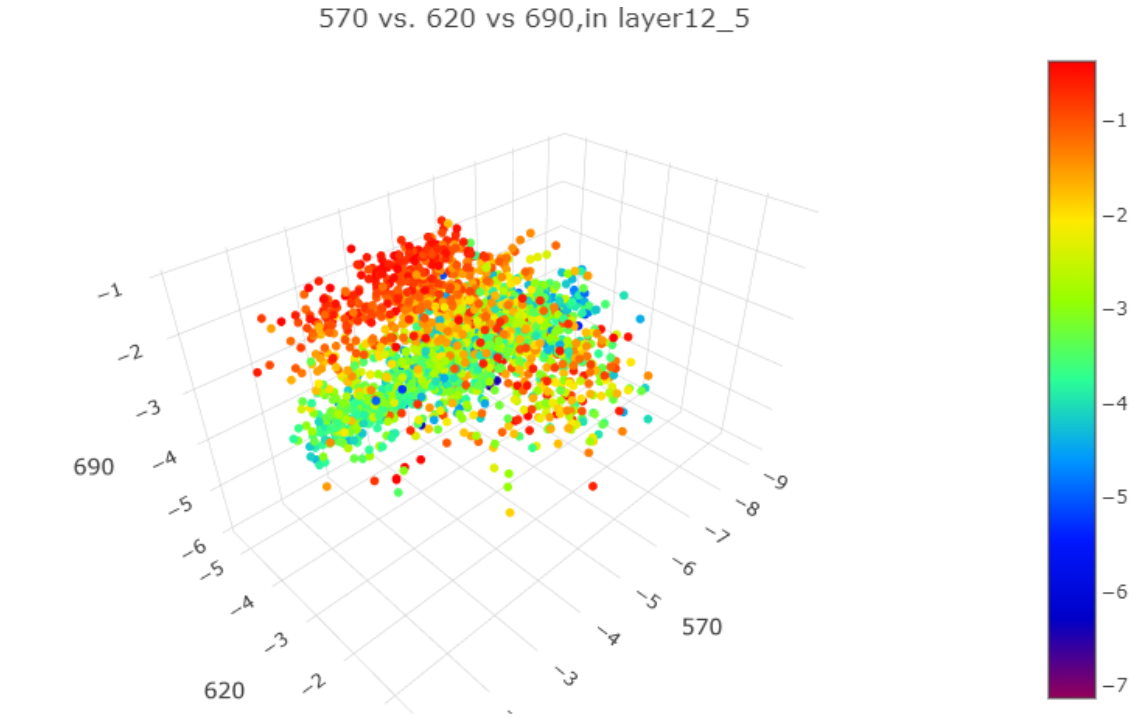


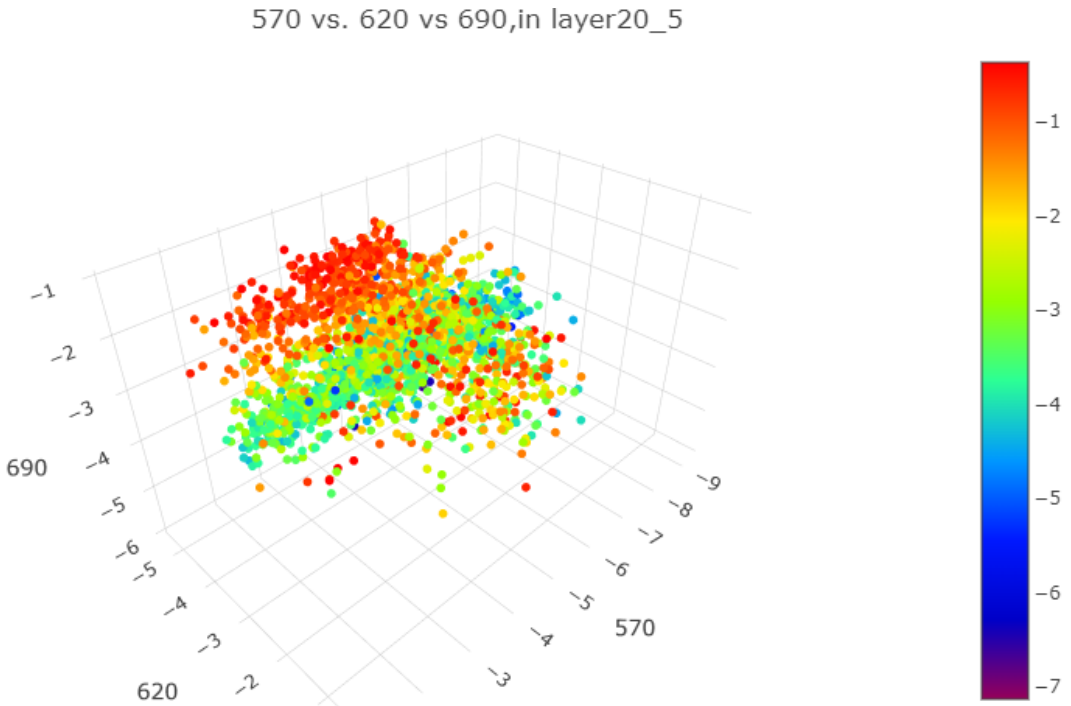
For the figures above, the intensity of color serves as an indicator of the strength of correlation, with darker shades representing greater correlation. Notably, if a line connects two variables, it signifies that these two variables exhibit the highest correlation compared to any others. Examining both layer12_4 and layer12_5, a notable line connects Opal_690 and Opal_520, providing empirical support for our hypothesis regarding their high correlation. Contrary to this, as previously mentioned, in layer20_5, the correlation between Opal_690 and Opal_520 is not particularly significant; however, it is intriguing to observe a connecting line between Opal_690 and Distance, indicating a noteworthy correlation between these variables. This unexpected correlation in layer20_5 adds a layer of complexity to our understanding, highlighting the importance of considering layer-specific dynamics in interpreting correlation patterns.

3D Plot

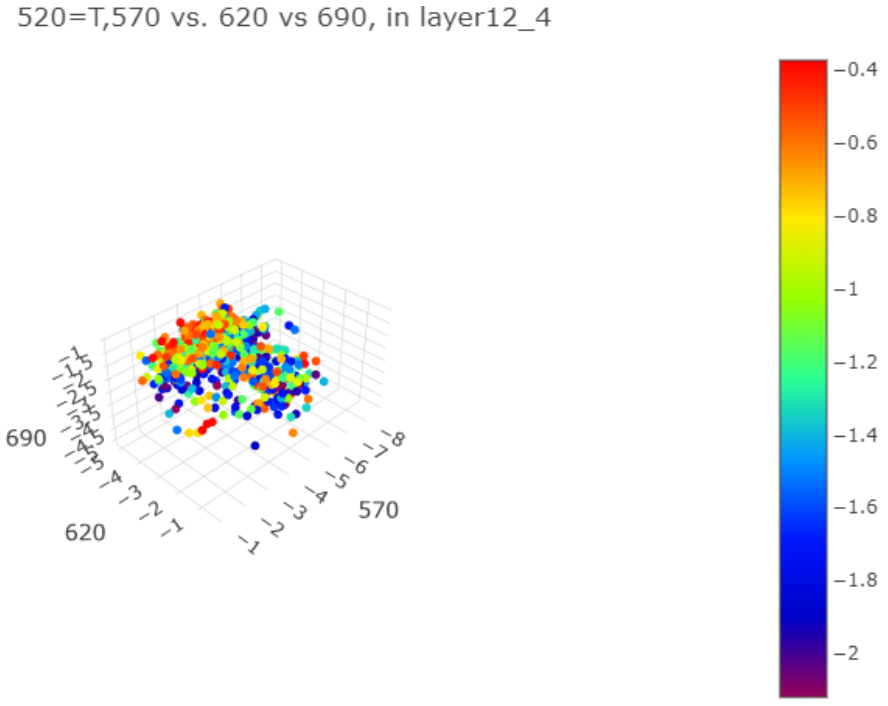
Since there is more than two variables we're dealing with, an idea of visualization in 3-D version is inspired by our advisor. But knitting the moving pictures of 3D plots is not available, besides the screenshots provided with only one or two perspectives, we will give you our original codes so that you can reproduce what we got so far. And please be aware that the scales labeled below are all after log-transformation aimed for a more clear view.

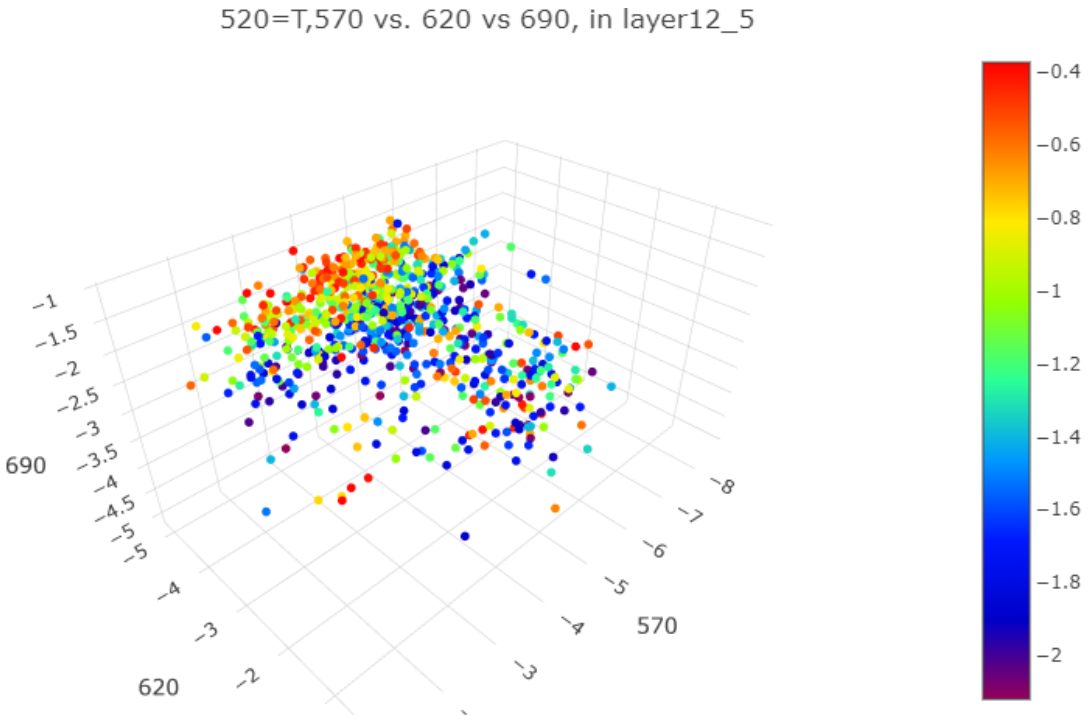

570 vs. 620 vs 690,in layer12_4
−3
−4
−5
−6
−5
−4
570
−5
−6
−7
−8
−9
1
0
−1
−2
−3
−4
−5
−6
−7

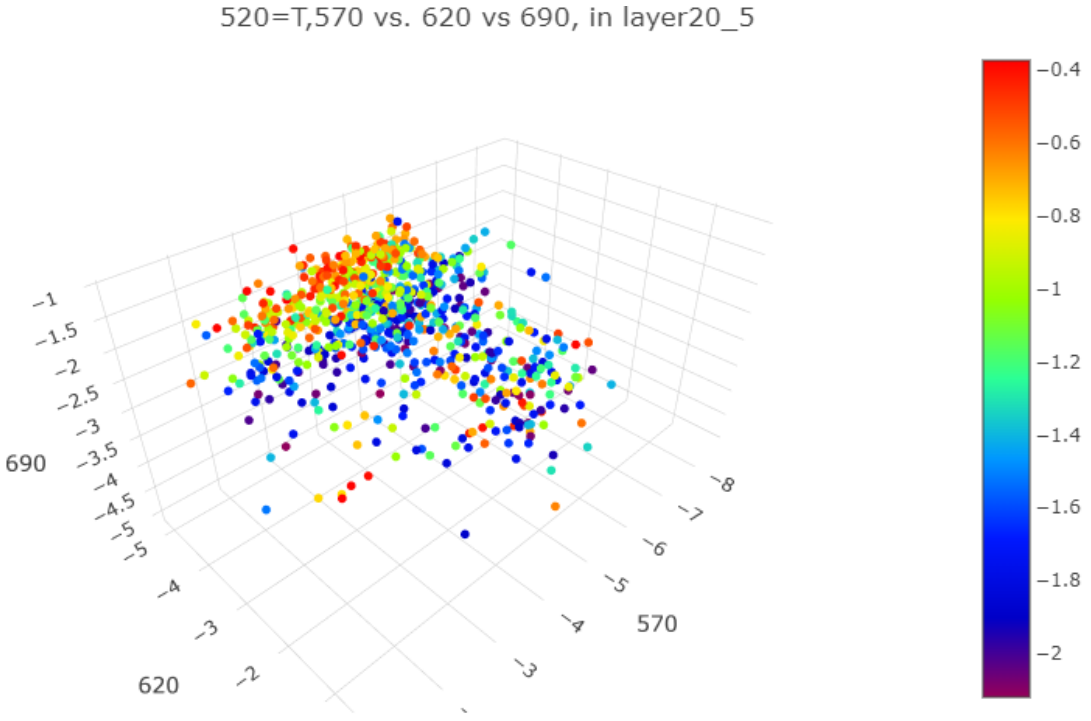

570 vs. 620 vs 690,in layer12_5
−1
−2
−3
690
−4
−5
−6
−5
−4
−3
620
−2
−3
−4
−5
−6
−7
−8
−9
570
−1
−2
−3
−4
−5
−6
−7



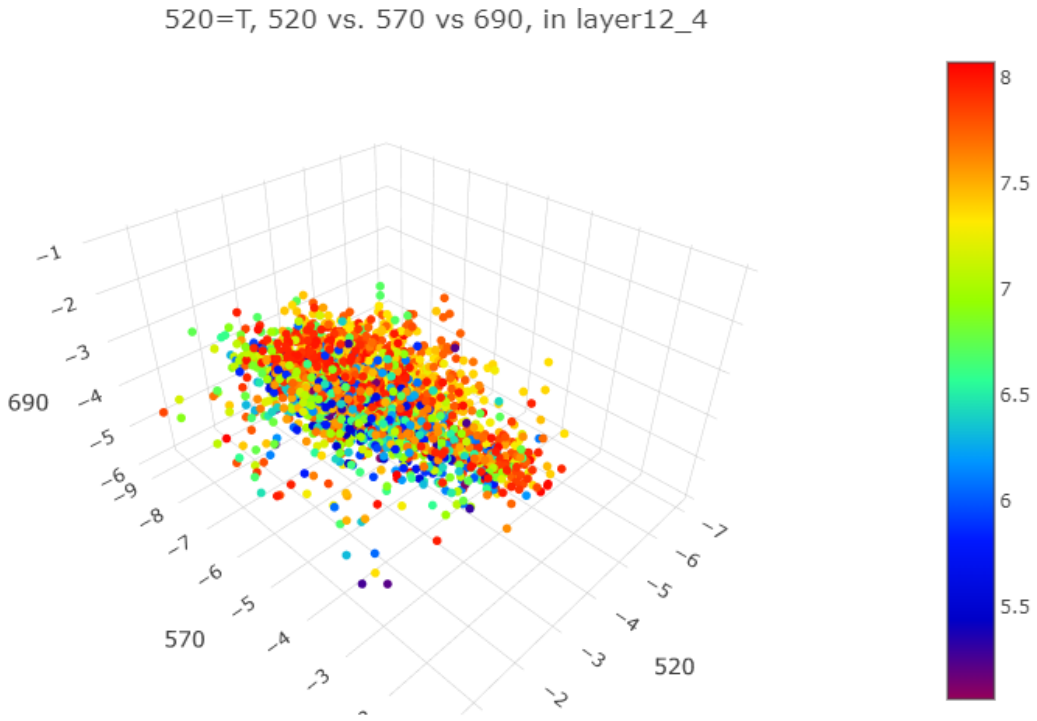
The plots above show the relationship among each gene in terms of intensity level of 520. But these plots accounts all the cells in the sample no matter there is 520 included or not. So it's hard to observe significant relationship among different genes. Next we decided to only include the cells with 520 detected.

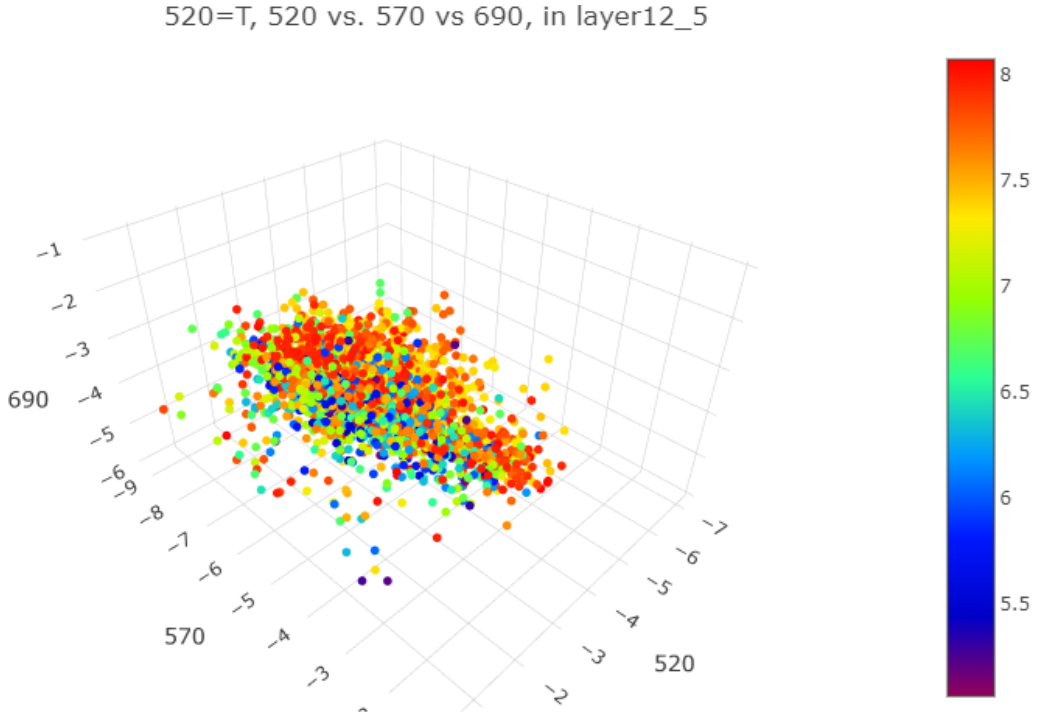


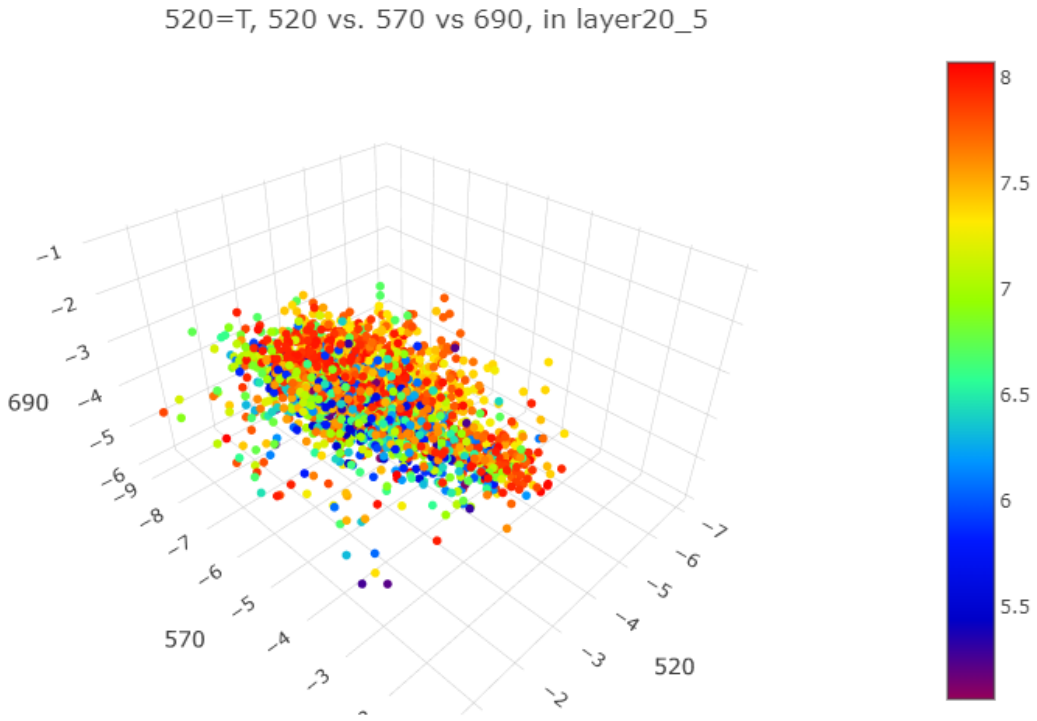




As removing the cells without 520, some relationships could be observed based on the 3D plots. For example, in Layer12_5, As the intensity of 520 raise, the intensity of 690 increases while 570 seems to show a reverse relationship with 520 compared to 690. Also, 620 doesn't have obvious relationship with other genes. So for further investigation, we decided to remove 620 which is consider meaningless for showing no any trends with the others in these three plots. Instead 520 is included below.


520=T, 520 vs. 570 vs 690, in layer12_4
690
570
520
8
7.5
7
6.5
6
5.5


520=T, 520 vs. 570 vs 690, in layer12_5
690
570
520
8
7.5
7
6.5
6
5.5



To get 520 involved, the intensity level as colored option is no longer appropriate. Besides intensity level, we thought there might also be some contribution made by distance to the distribution of cells in EC region. So instead of intensity level, we redo the previous process but colored in terms of distance. And we found that In layer12_5 as the intensity of 520 decreases, the intensity of 690 also decreases. And as intensity of 520 held constant, the higher the intensity of 690 is the more distant from the edge of EC.

Here are only couple examples from each section of plots. We hope these plots would help our client to seek more essential information for their study.

Conclusion

The project's application of robust statistical methods revealed intricate patterns in gene expression. Notably, a significant correlation between Opal_690 and Opal_520 was found, especially in layers 1 and 2 of the EC. These results highlight the complex interplay of gene expression, which varies distinctly across different layers. Furthermore, the spatial analysis of gene distribution underscored how gene expression intensity is related to proximity to the EC edge, adding a spatial dimension to our understanding of gene distribution.

The key to the successful output of this consulting project is the teamwork of our group. Everyone paid attention to what we're assigned, brainstormed any idea which would be the approach to solve the issues, help generate different plots for the clients to have a better visualization of her interest. The most important statistical skills involved would be the R programming. Most of the time was spent on Rstudio trying to figure out different codes for visualization. Also members learnt some plotting skills which they never knew before during the coding.

In summary, we have not only came up with some useful results for the clients but also advanced our understanding of what we learnt in class. Hope our findings in the Entorhinal Cortex would give insight and inspire more idea and even breakthroughs in neurology.