Consulting Project

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

2023-12-14

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 divide the three layers into two parts, one is the data that all cells contain Opal_520, the other dataset have all cells whether it contains Opal_520 or not. Most of the time, we use the data with all cells contain Opal_520. Here is an example of head 5 rows of the data.

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 another dataset, we check the existence of genes in the cell and add four columns with boolean output. We just use this dataset with 3D plot. Here is an example of head 5 rows of the data.

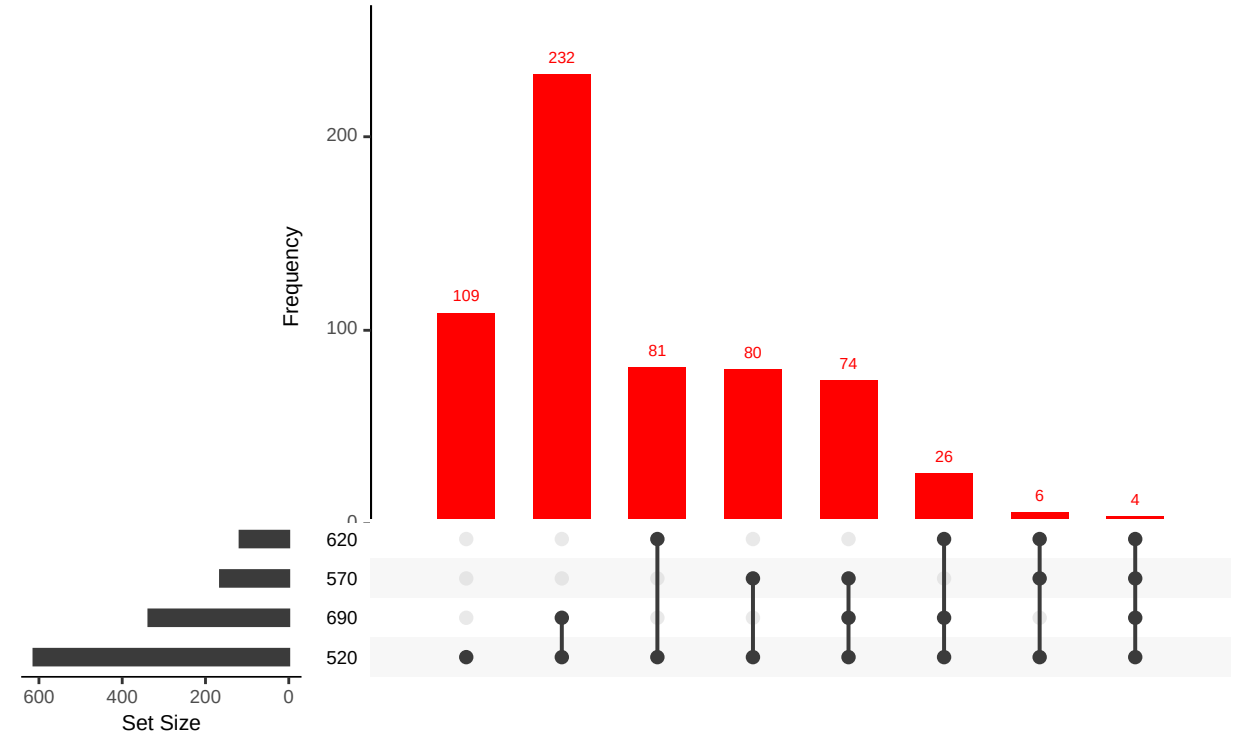
MFI520	MFI570	MFI620	MFI690	dist	IND520	IND570	IND620	IND690
0.3483	0.1596	0.0225	0.1164	2871.8301	TRUE	TRUE	FALSE	TRUE
0.2152	0.1041	0.0196	0.1136	2866.8936	TRUE	TRUE	FALSE	TRUE
0.5518	0.0258	0.016	0.2296	2861.261	TRUE	TRUE	FALSE	TRUE
0.4816	0.0202	0.02	0.3153	2868.6372	TRUE	FALSE	FALSE	TRUE
0.2459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

Both of the data clean the value with fluorescent intensity equals 0.

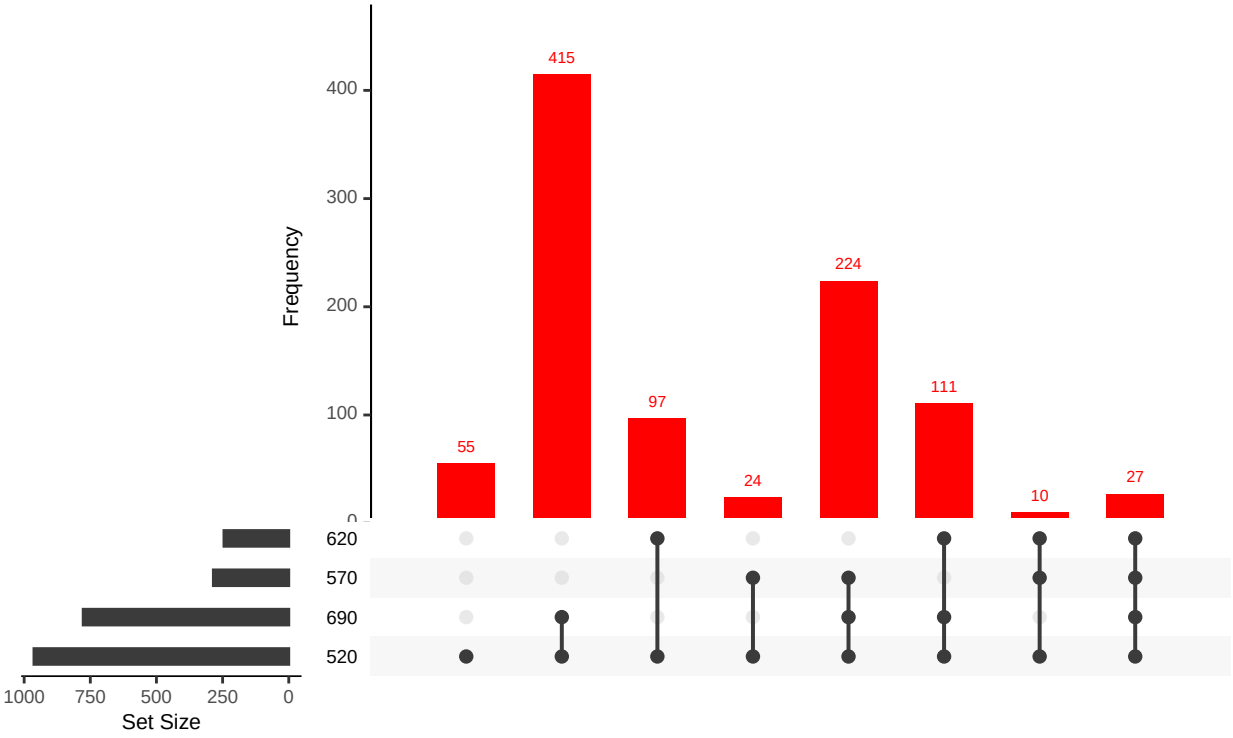
Visualization

Upset Plot

We use the Upset plot to see the distribution for three layers.

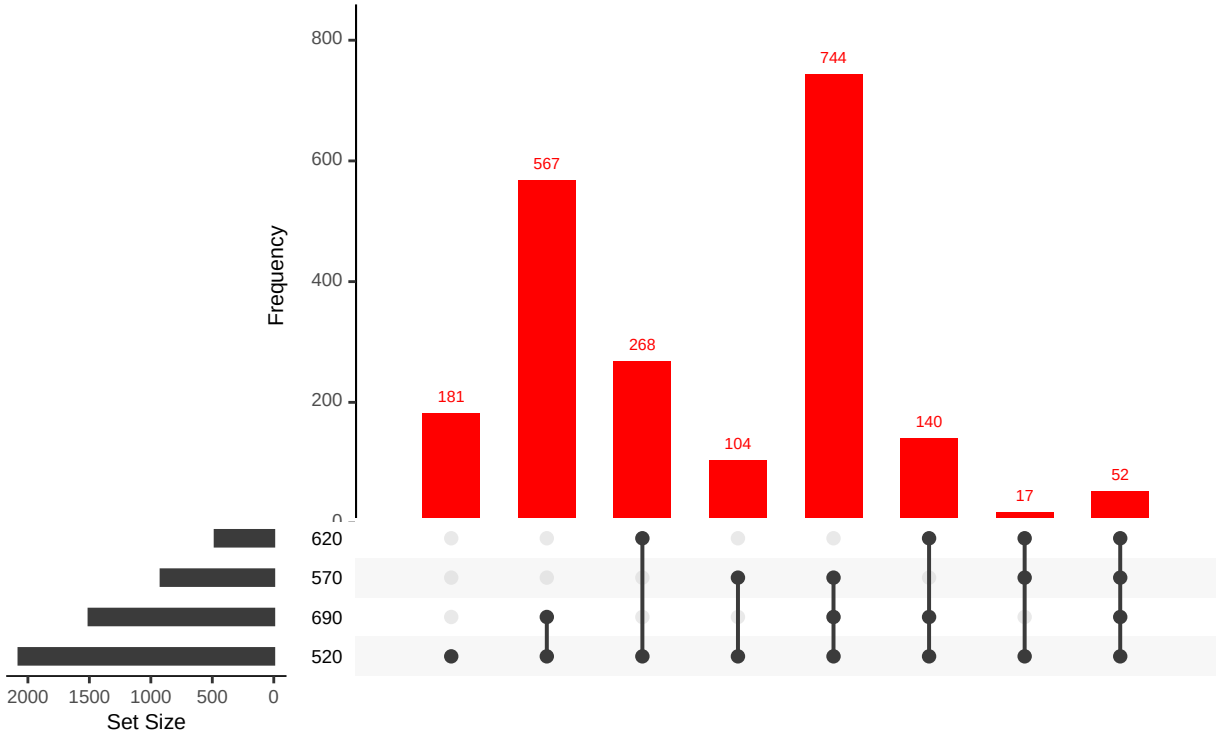


The figure above shows the distribution for layer20_5. We use the data all cells contain Opal_520, so the number of Opal_520 implies the number of cells in the data which is around 600. And we can see the cell with 520:690 has the highest frequency with the number 252 in the layer.



The figure above shows the distribution for layer12_4. We use the data all cells contain Opal_520, so the

number of Opal_520 implies the number of cells in the data which is around 1000. And we can see the cell with 520:690 has the highest frequency with the number 415 in the layer.

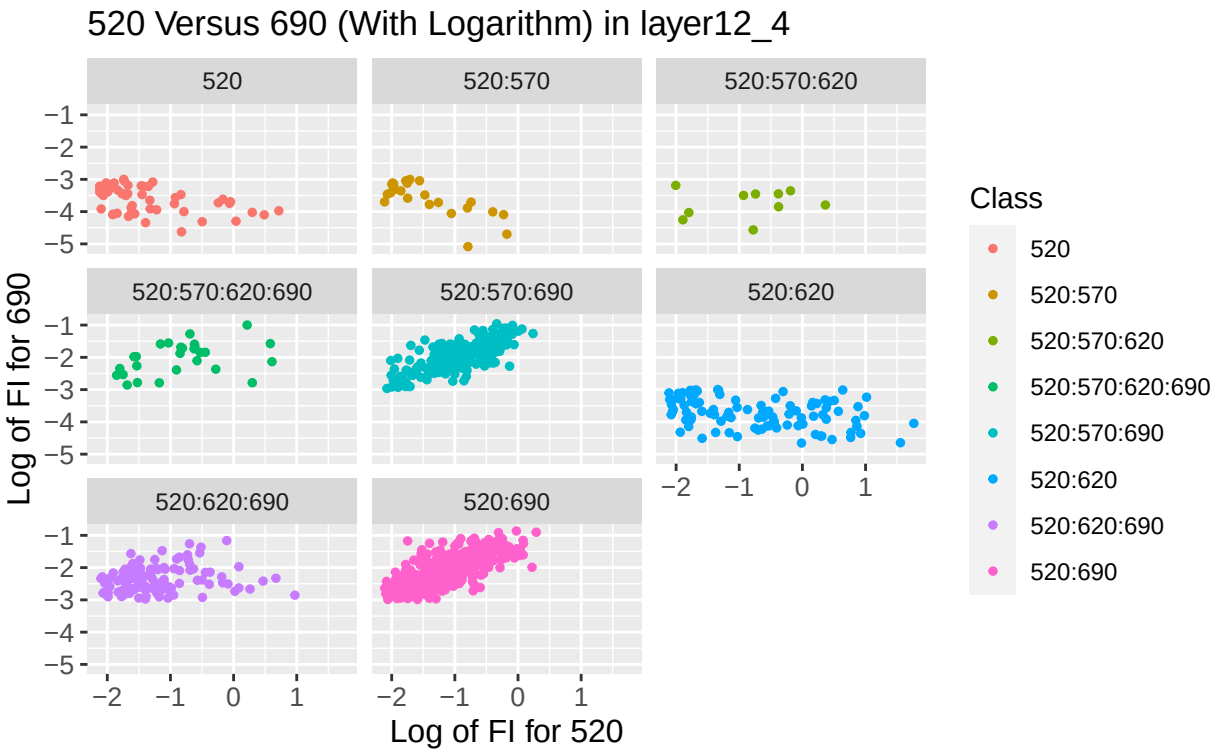


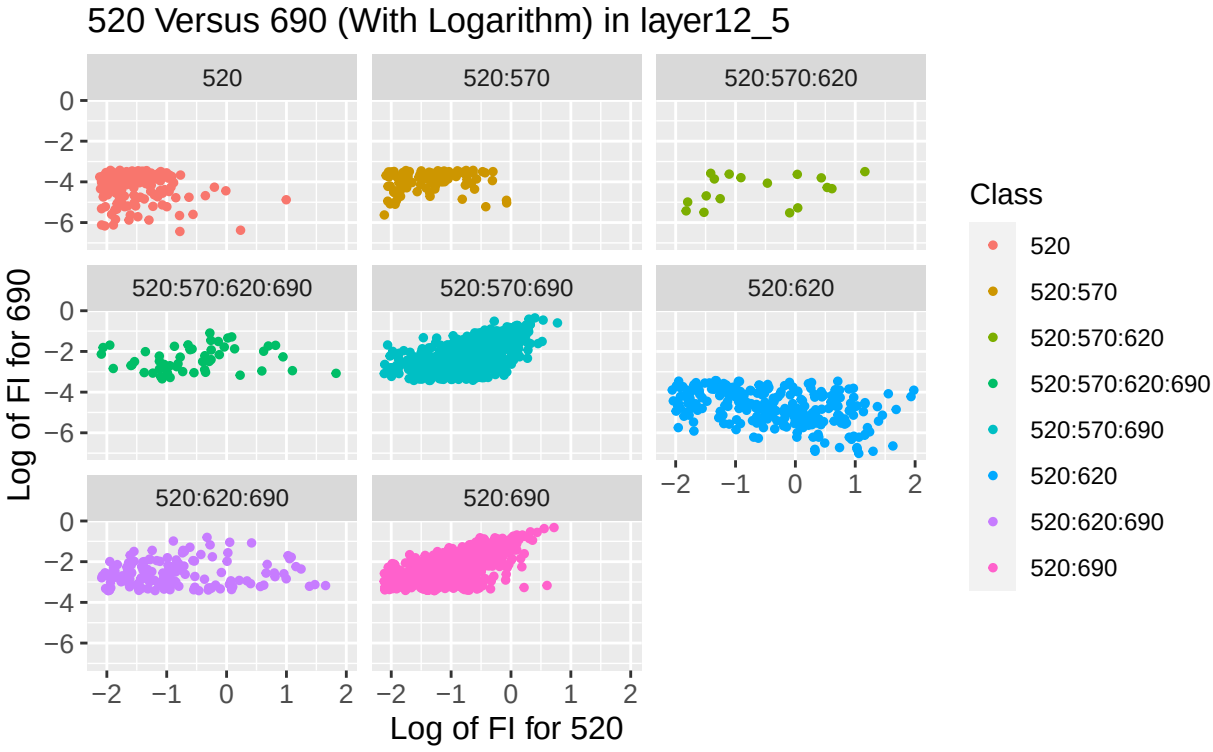
The figure above shows the distribution for layer12_5. We use the data all cells contain Opal_520, so the number of Opal_520 implies the number of cells in the data which is around 2000. And we can see the cell with 520:570:690 has the highest frequency with the number 744 in the layer.

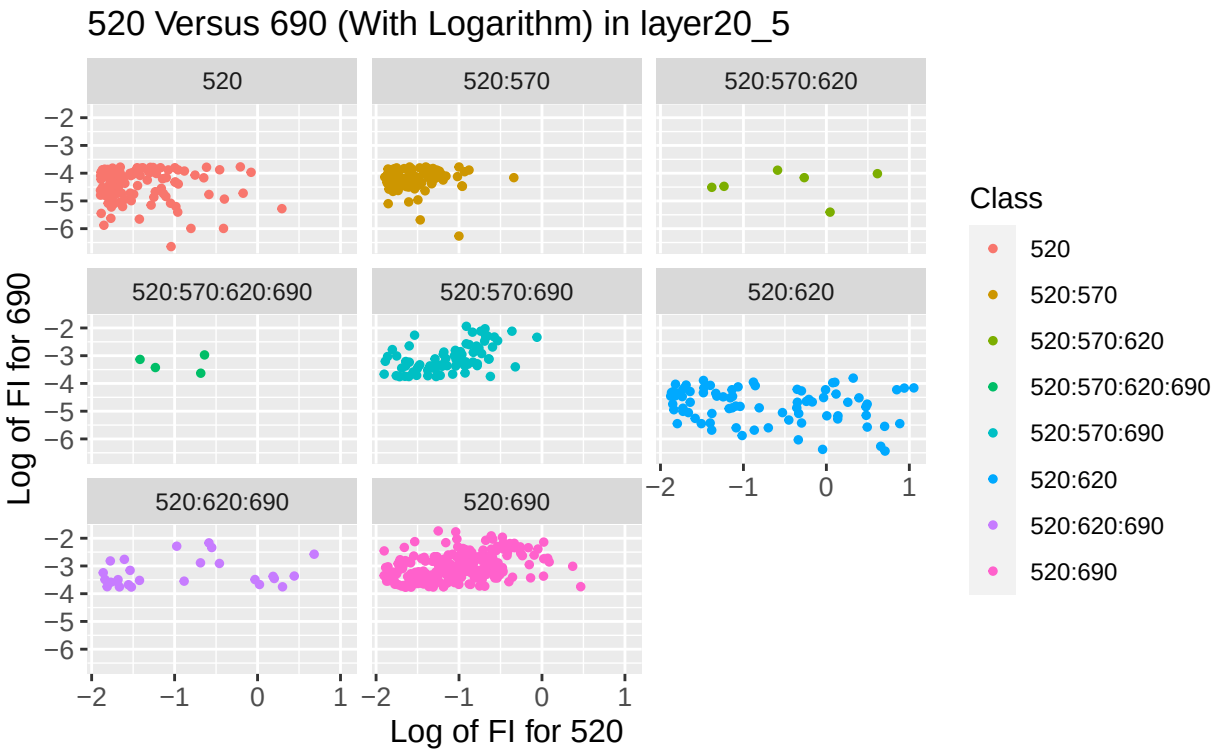
As a summary with this three figures, we can see that Opal_690 appears more than Opal_570 and Opal_620. In layer20_5 and layer12_4, we both have 520:690 with the highest frequency in the layer. But in layer12_5, though 520:690 has a big frequency value, 520:570:690 has the biggest frequency value in this layer. Interestingly, Opal_520 and Opal_690 appears very often. So in the next step, we want to see the relationship between Opal_690 and Opal_520.

Correlation

We use the Scatterplot to see the distribution of Opal_520 and Opal_690. In the plot, the x axis shows the log value of the fluorescent intensity for Opal_520, the y axis shows the log value of the fluorescent intensity for Opal_690.

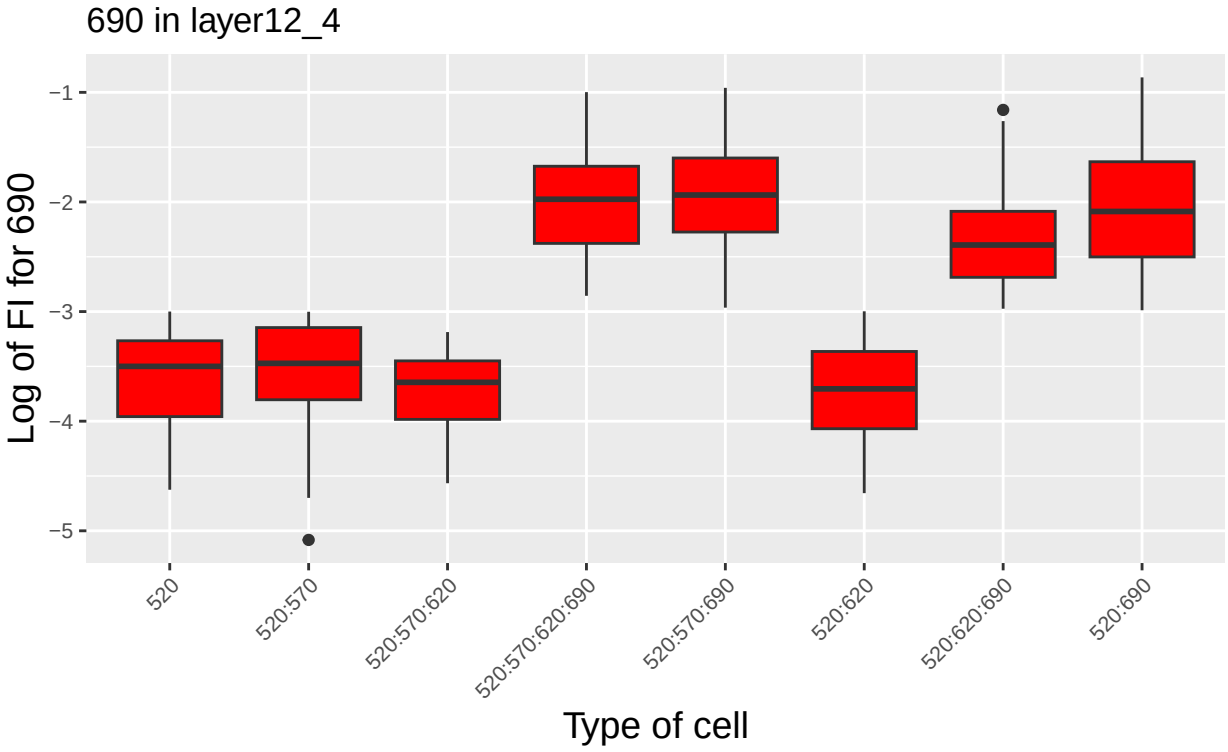

520 Versus 690 (With Logarithm) in layer12_4
520
520:570
520:570:620
520:570:620:690
520:570:690
520:620
520:620:690
520:690
Log of FI for 690
Log of FI for 520
Class
520
520:570
520:570:620
520:570:620:690
520:570:690
520:620
520:620:690
520:690

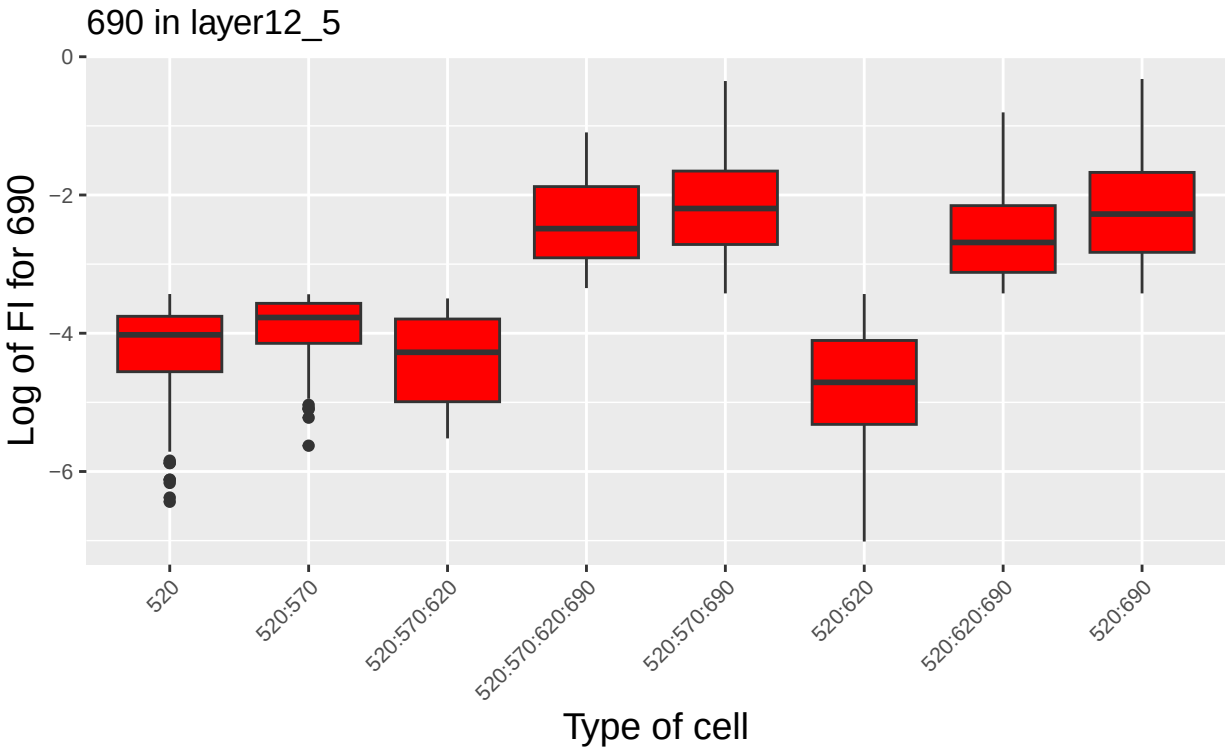

520 Versus 690 (With Logarithm) in layer12_5
520
520:570
520:570:620
520:570:620:690
520:570:690
520:620
520:620:690
520:690
Log of FI for 690
Log of FI for 520
Class
520
520:570
520:570:620
520:570:620:690
520:570:690
520:620
520:620:690
520:690

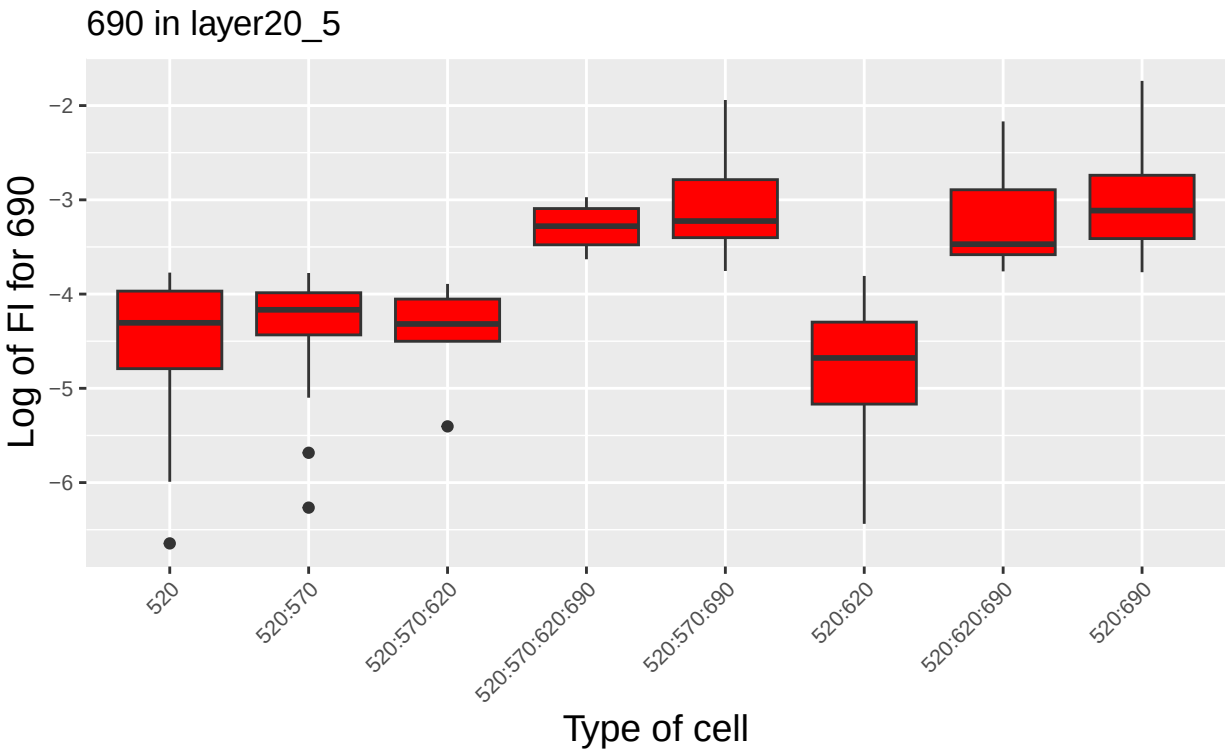


For the figures above, we can see the cells contain Opal_690 will have a high value with Opal_690 which makes sense. In layer12_4 and layer12_5, the cell 520:570:690 and 520:690 shows a positive relationship for Opal_520 and Opal_690. When the value of Opal_520 increases, the value of Opal_690 also increases. In layer20_5, we can not find significant correlation between Opal_520 and Opal _690. We are not sure if this correlation is common for all the layers or only happens in layer12_4 and layer12_5, since there is a slight positive correlation in layer20_5 but not as significant as layer12_4 and layer12_5.

Boxplot


690 in layer12_4
Log of FI for 690
−1
−2
−3
−4
−5
520
520:570
520:570:620
520:570:620:690
520:570:690
520:620
520:620:690
520:690
Type of cell

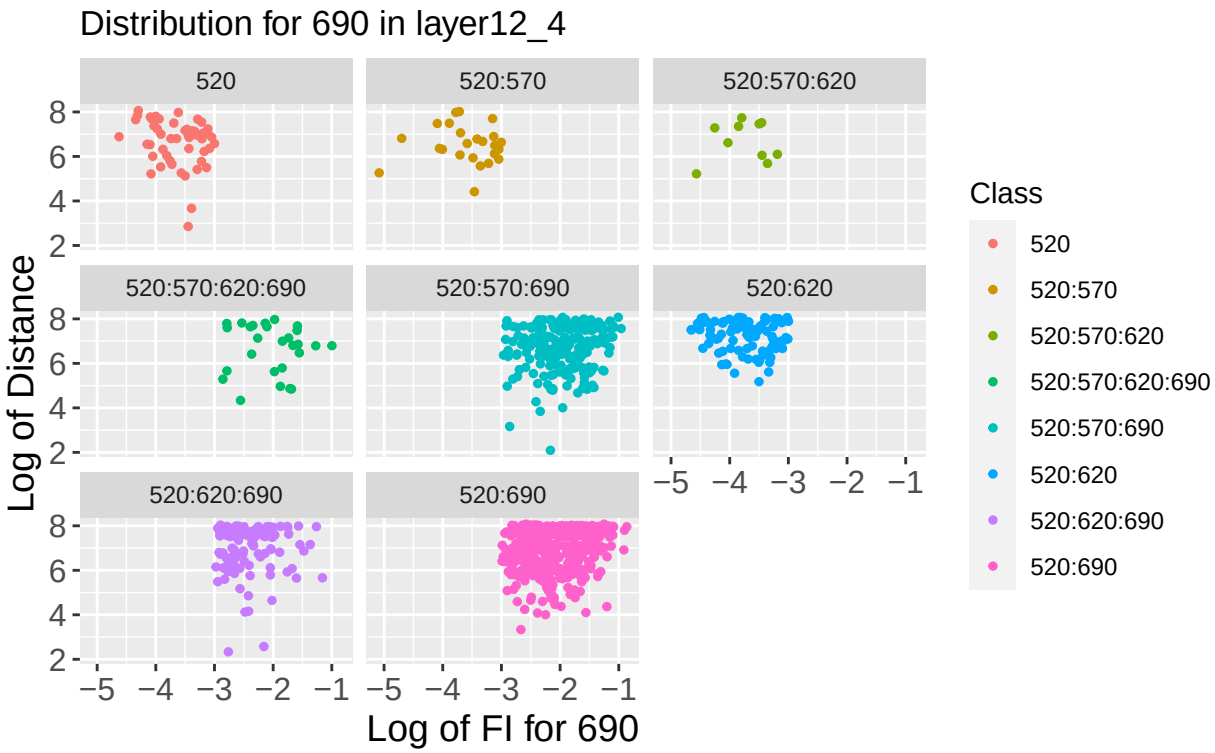

690 in layer12_5
Log of FI for 690
0
−2
−4
−6
520
520:570
520:570:620
520:570:620:690
520:570:690
520:620
520:620:690
520:690
Type of cell

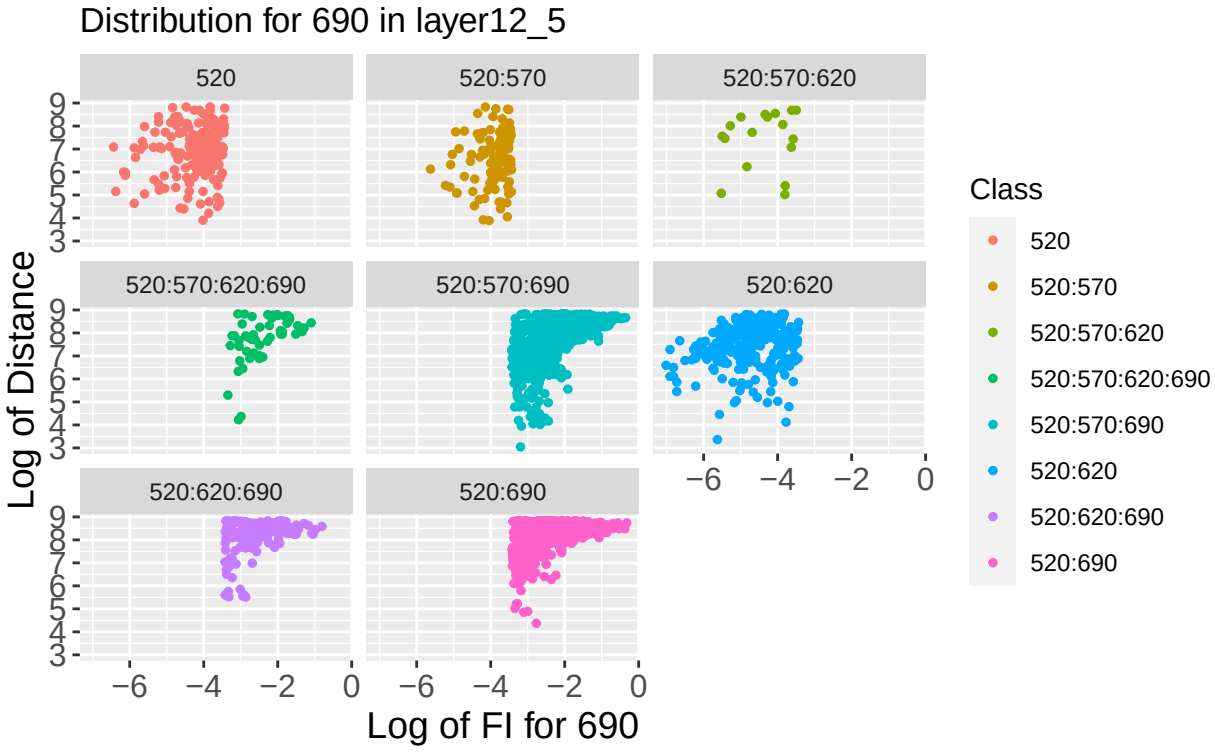


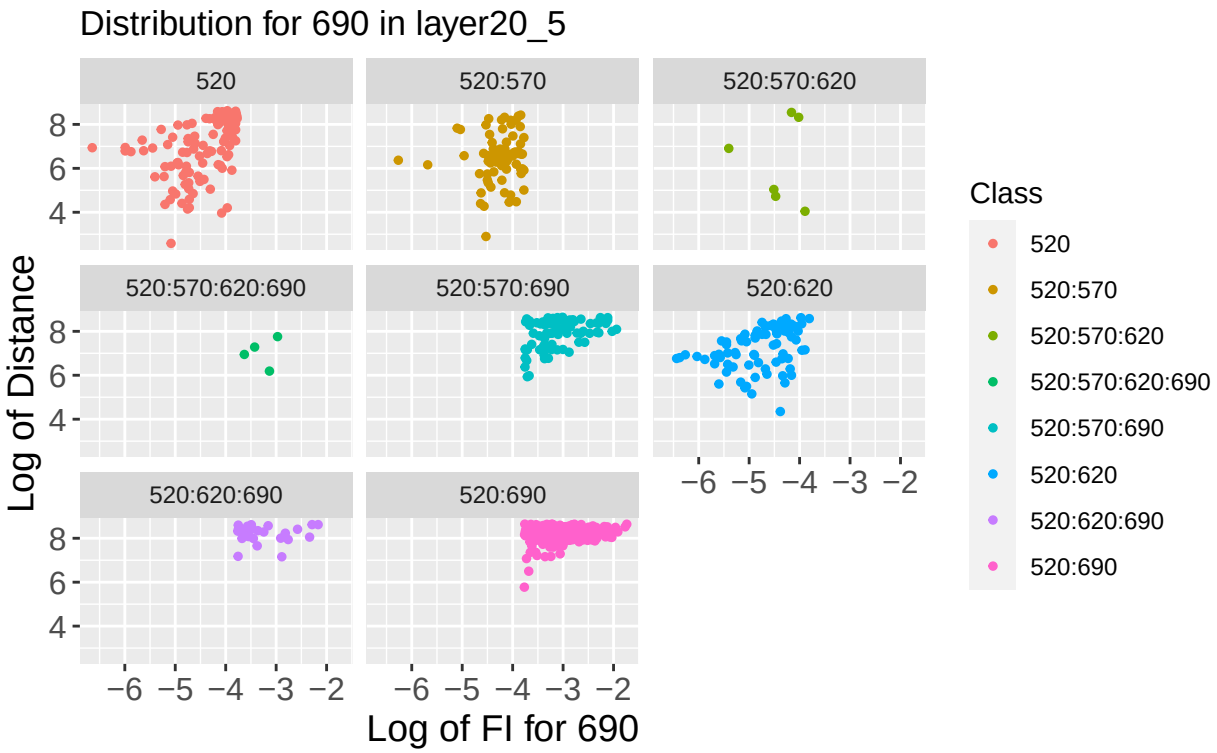
For the figures above, we can see the cells contain Opal_690 will have significant higher value than the cells don't contain Opal_690.

Distribution of Distance and Opal_690

Since Opal_690 has some correlation with Opal_520, we want to find if there is any specific correlation for Opal_690 and Distance. In the plot, the x axis shows the log value of the fluorescent intensity for Opal_690, the y axis shows the log value of the Distance.


Distribution for 690 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 Distance
Log of FI for 690
Class
520
520:570
520:570:620
520:570:620:690
520:570:690
520:620
520:620:690
520:690


Distribution for 690 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 Distance
Log of FI for 690
Class
520
520:570
520:570:620
520:570:620:690
520:570:690
520:620
520:620:690
520:690



For the figures above, we focus on the cells with 520:690, 520:570:690, 520:620:690, and 520:57:620:690. It is most significant for layer12_5 that when the value of Opal_690 is small, the distance has a big range from low to high. But when the value of Opal_690 becomes bigger, we can see that the distance shrinks to a low range wiht only big value. For layer20_5, we can see that it has a small range with big value of distance whether the value of Opal_690 is big or small.