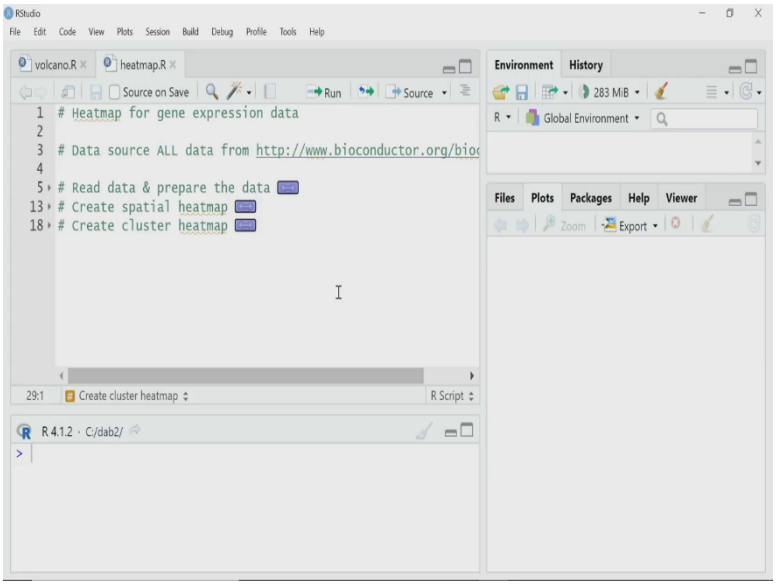
Data Analysis for Biologist
Professor Biplab Bose
Department of Biosciences and Bioengineering
Mehta Family School of Data Science And Artificial Intelligence
Indian Institute of Technology, Guwahati
Lecture 22
Heatmap and Volcano Plot

Hello everyone, some of the experiments in biology are high throughput experiments that generate a higher dimensional data. For example, take the example of microarray or RNA Seq. In these experiments, you are collecting the expression data of hundreds and thousands of gene from 1, 10, 100 of samples in one experiment.

Now, this is higher dimensional data, the data set itself is a large one and it is a challenge to visualize this whole data set comprehensively. In this lecture, we will discuss about two method to comprehensively visualize a high throughput high dimensional data particularly data coming from experiment like RNA Seq and microarray. These two techniques are plot of heatmap and volcano plot. So, I will start with heat map.

(Refer Slide Time: 01:35)

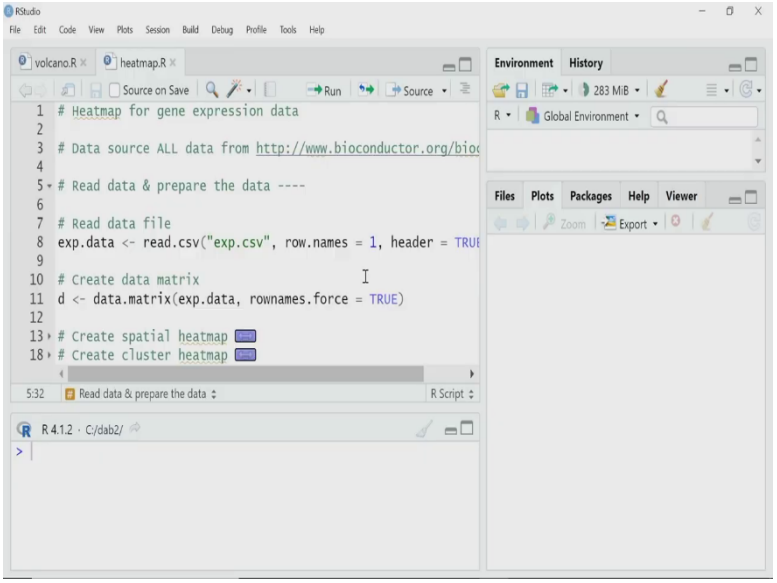


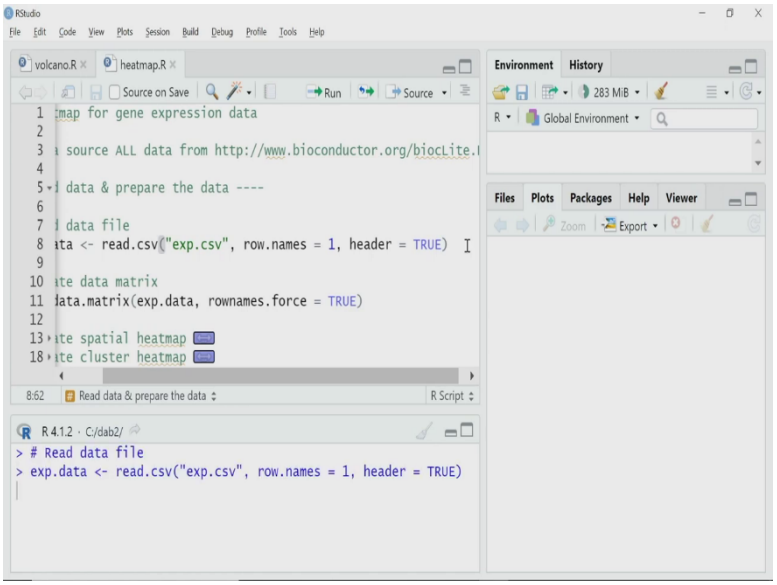
I will explain heat map and how to draw it using are using R, a script that I already have written. What I will do, I will first discuss the data and then I will directly plot the diagram, the heat map, and then I will explain what we communicate through heat map. And then we will go into nitty gritty and details of how you draw that particular type of figure or plot.

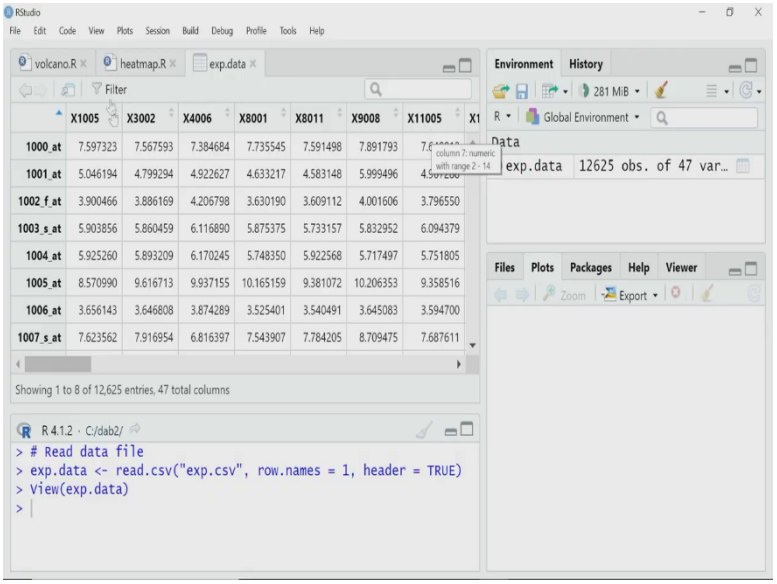
So, the data set that I am using, it is a data set coming from a microarray experiment of a particular type of leukaemia, ALL particular B cell and T cell ALL and I have downloaded the whole data set. From that data set, I have taken a part of it only a small part of it with I think 47 or 48 sample data are there, but 1000 of features for each of the samples are there.

And these are for ALL with specific type of type of mutations. Now, we do not need to go into details of that biology. We have to understand that this is a higher dimensional data where I have 1000 of microarray feature. And for each sample, we are measuring those features. So, let me first load that data and look into what the data we have.

(Refer Slide Time: 02:49)







exp.data ← read.csv("exp.csv", row.names = 1, header=TRUE)

d ← data.matrix(exp.data, rownames.force = TRUE)

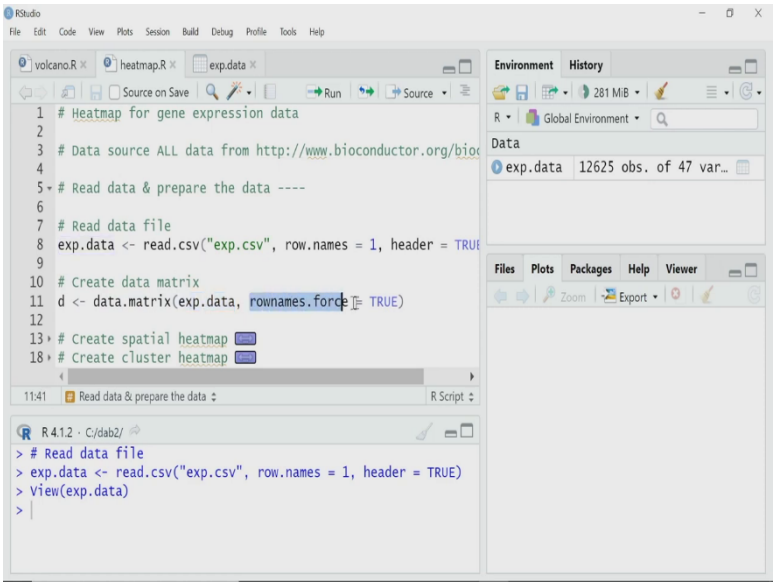
So, I have created the data in a CSV file. So, I am reading it using read CSV, I am keeping the header true. So, and the row names, the first column of the data set has the name of the features or in a way the genes right, so that I have specified that these are the row names, and I am saying the row names equal to one.

So, let me read the data first. So, in the environment tab, let me open it, it may take a few seconds. Because the very large data set because you can see it has 12,625 observations for 47 variables. So, what I have, you can see the columns, I have X1005, X3002 these are samples, and I have 47 samples, 47 patient samples.

For each sample, we have microarray feature data. And these row names, like for example, 1000, underscore 80 is a possibly if I remember correctly, is the feature for a particular type of category. So, these each row is for one of those features in the microarray. So, in a way, they represent some gene.

And you can imagine we have 12,625 such features. Now, how can I visualize this data comprehensively, and to do that actually, heat maps are very useful. So, what I will do, I will draw the heat map and I will take then explain the different elements of heat map.

(Refer Slide Time: 04:20)



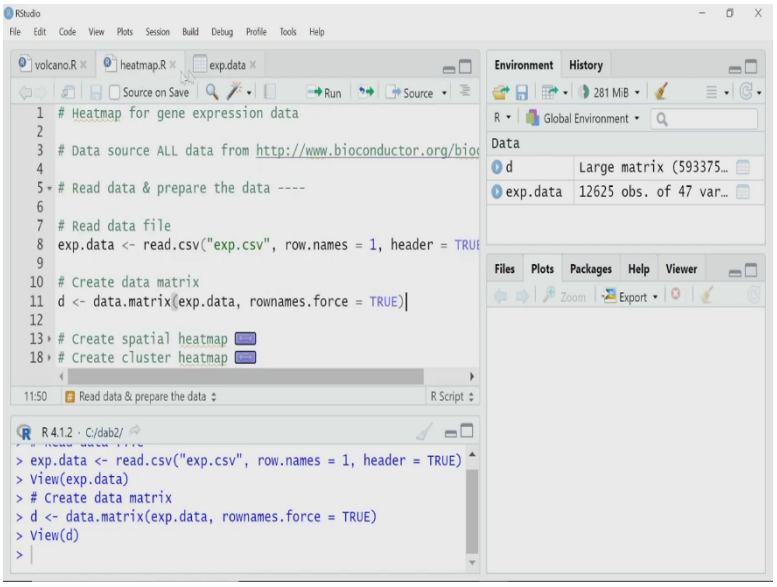
exp.data ← read.csv(“exp.csv”, row.names = 1, header=TRUE)

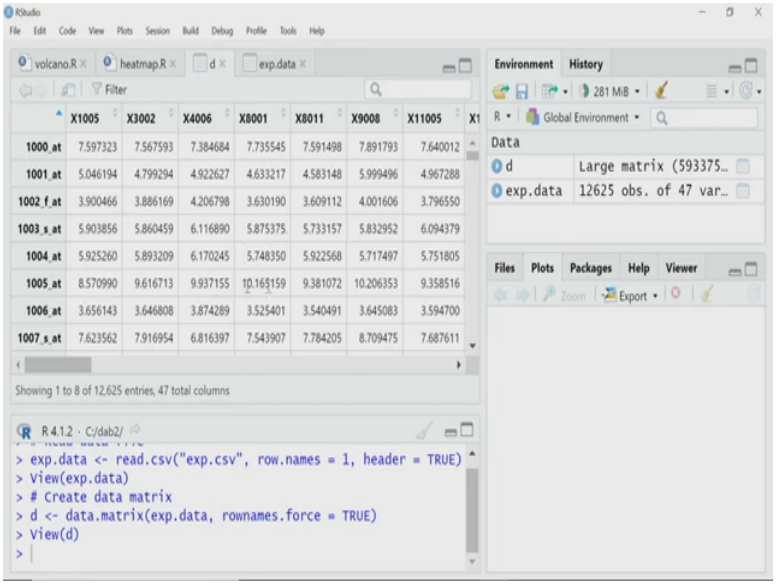
d ← data.matrix(exp.data, rownames.force = TRUE)

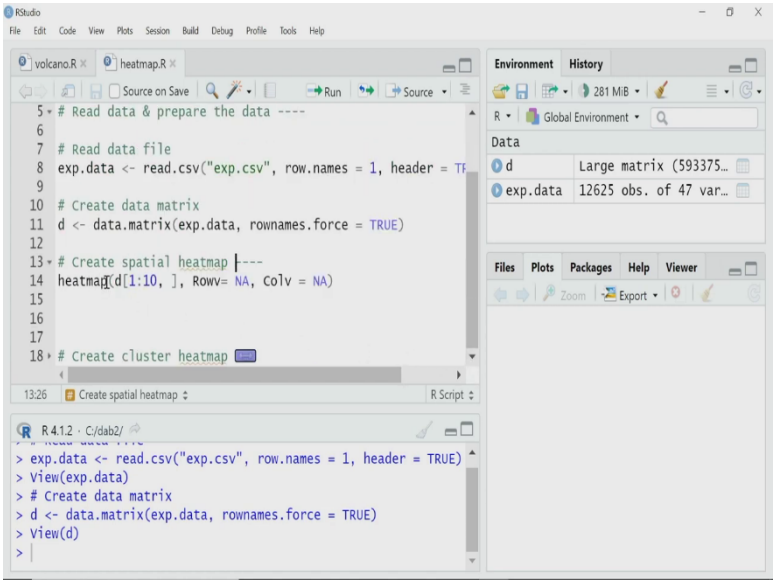
Now, to draw the heat map, I will use the inbuilt default heat map function that comes with R. And that Heatmap function takes a matrix as an input or an argument, so data should be in matrix format. Right now the data that I have loaded is in data frame. So, I will convert this data frame into a matrix.

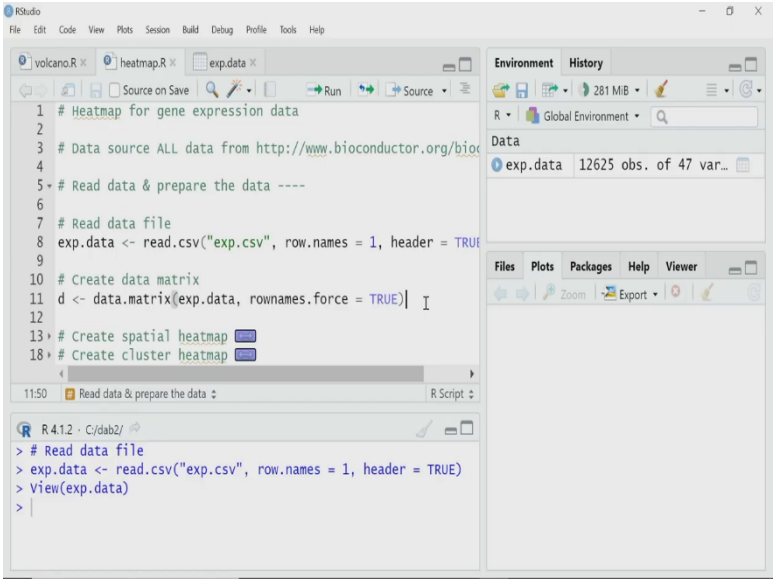
So, that is what I do in this line where I use data dot matrix this function data dot matrix, and I give experiment exp dot data that is the data frame as one argument and then I say that the row names are true, you force the row name to be true, so that then the matrix also, the row names are retained. That means I want to retain the row names present in the data frame. And remember those row names are nothing but the feature names. So, I want to retain that.

(Refer Slide Time: 05:11)









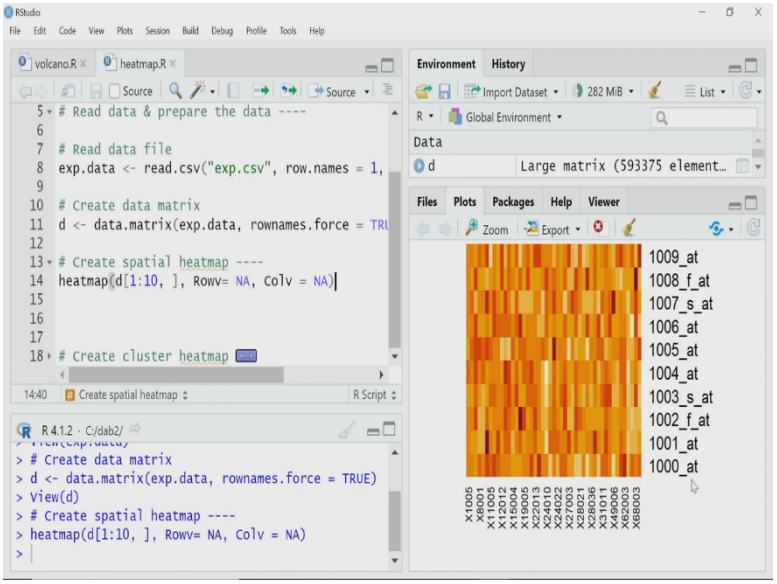
heatmap(d[1:10,], Rowv = NA, Colv = NA)

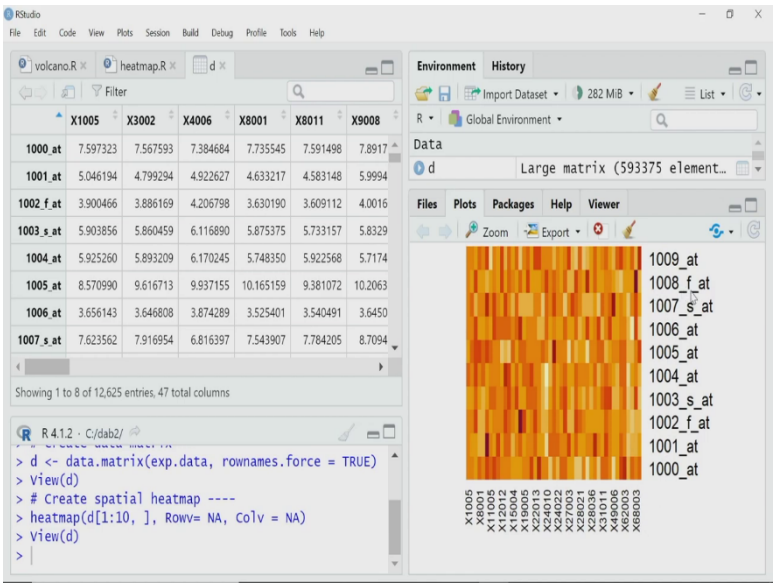
So, let me I will convert it into a matrix. And now I have a large matrix you can see here in environment pane, if I double click, you can see that data in the matrix, remember the header is also there and the row names are also there. So now, I will draw the heatmap. And to draw the heatmap, I will use the default heatmap function. The first argument for heat map function is the data itself. Now D is my data that is a matrix.

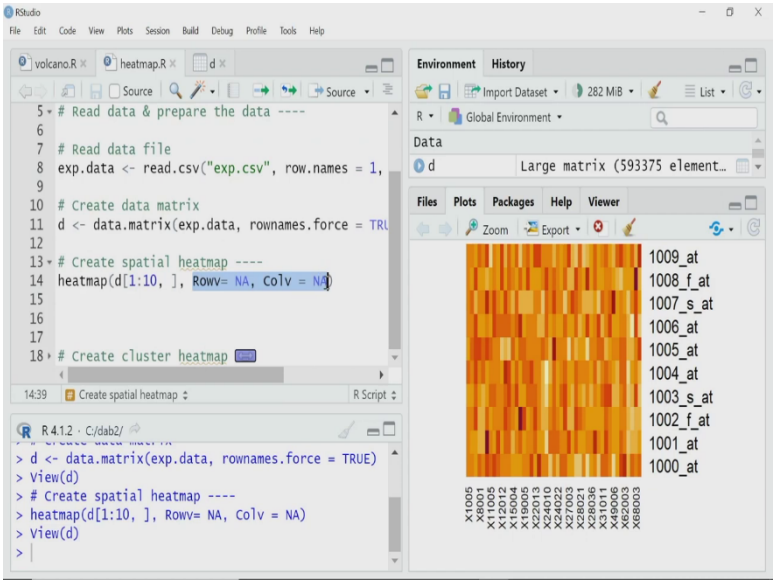
Just to keep a clarity here, I do not want to create a heat map for all those 12,000 features. I am taking only the first 10 rows, that means I am taking only 10 features, so that the diagram should be very clean and I will be able to explain the diagram easily to you. Otherwise, I would have may have taken the whole data set D.

So, what I am taking I am saying, okay take the 1 to 10 row for D and I am leaving this columns space empty that means all the columns and then there are some arguments, I will explain this argument once I create the diagram, I will create the diagram, let me expand the space for plot tab, so that we can visualize the diagram very clearly.

(Refer Slide Time: 06:34)







heatmap(d[1:10,], Rowv = NA, Colv = NA)

So, here is my heatmap what you can see I have the features or gene these are, these 1000 underscore at, 1001 underscore at these vertical axes they represent, each of these rows represent the feature. And these columns actually represent each of the sample and those samples number are written in the horizontal axis.

So, now, if you imagine that these rows, rows for feature and the column for a sample they intersect, so you have, the heatmap function has drawn a box there and in that box, it has filled the box with a colour and that colour has a code. So, the lower values in my data set. In my data set D if I have a lower value, that is colour coded by a faint colour, faint yellow.

Whereas the reading a particular value for a particular feature for a particular sample, if it is very high, then it is a very dark colour. So, by this colour code each of these rectangle created by the intersection of sample and the features, so, are filled, all of these rectangles are filled with a colour and the colour is telling me whether that that data for that feature in that sample is have a high rating or a low reading.

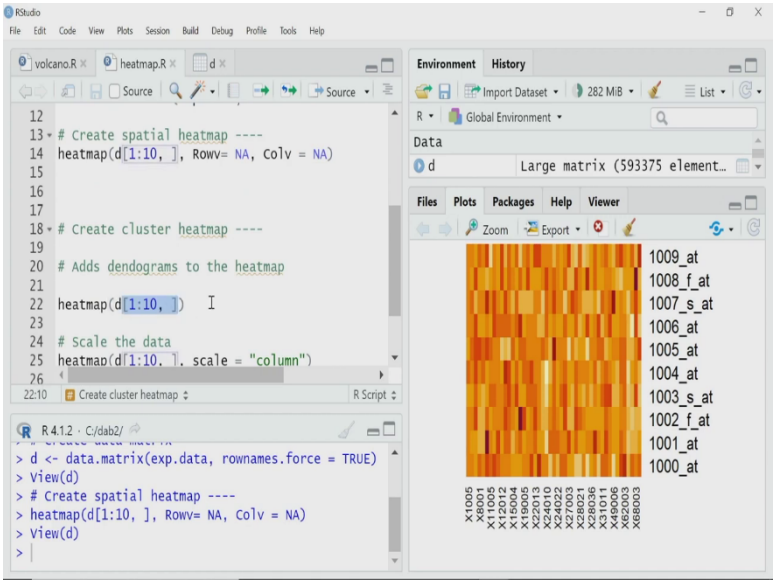
For example, if I look into here 1004 underscore at, row if you go in the middle I have a sample 24010 that is almost white that means that reading is very low. Whereas for the feature 1008 underscore f underscore at, the sample, here, the second sample from the right hand side has a very dark colour that means the reading for that feature in that sample is very high. This type of heatmap many a times is called spatial heatmap.

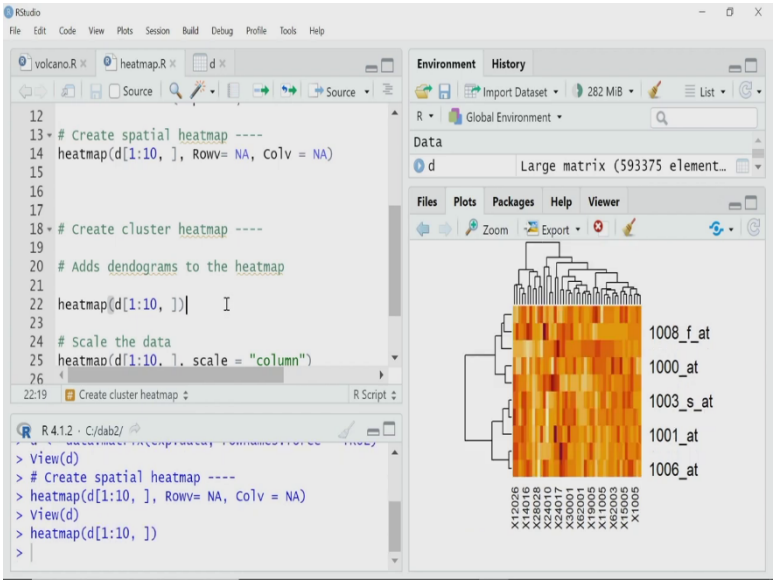
In this heat map, the data is arranged as I have provided in my data file in my data matrix. The heatmap function has not arranged it using some rule some other way. So, there is no manipulation done by the heatmap it is essentially as represented the numerical values of the reading by colour that is all. Now, many times you want to cluster these data each of these gene expression right into different cluster, each cluster will have similar behavior.

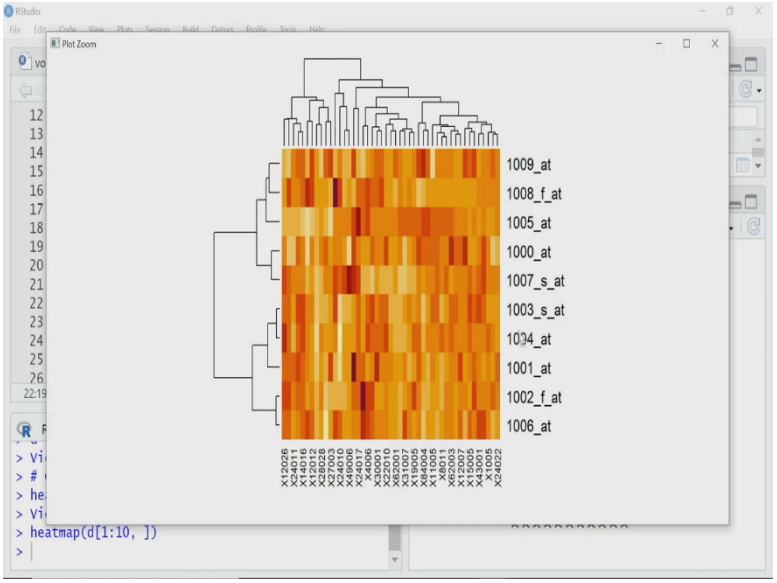
In that case, we add the clustering data along with this heatmap and we represent the cluster by a hierarchical dendogram. The heatmap function can also do that. That type of heatmap will be called cluster heat map. But in this case, I have not performed the clustering, I have not drawn the dendogram that is why when I call the heatmap function.

I have used two additional argument row v equal to NA and column v equal to NA. These two arguments told the heatmap function that I do not want any clustering of the data, just draw the spatial heat map without any clustering or the dendogram. Now, once I have discussed this clustering, it can be done by heatmap why do not we try to do that?

(Refer Slide Time: 09:59)







heatmap(d[1:10,])

So, I will move into the next section where I will now, will not provide these row v equal to NA and column v equal to NA. Rather, I will skip those two argument and use the same data set D, 1 to 10 rows, rest on all the 47 column, and I will call the heatmap function. So, here I call that.

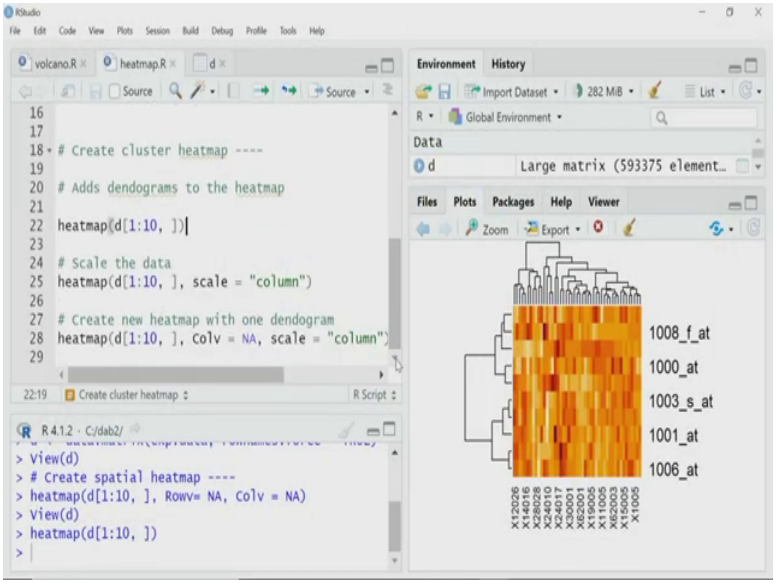
Now, you can see if I zoom out, I have dendrograms both on the horizontal axis, also in the vertical axis, the vertical axis dendogram here, which is Pars, this is actually clustering the genes and then drawing a hierarchical cluster. That means what this dendogram is saying that 1009 underscore AT and 1008 underscore F underscore at.

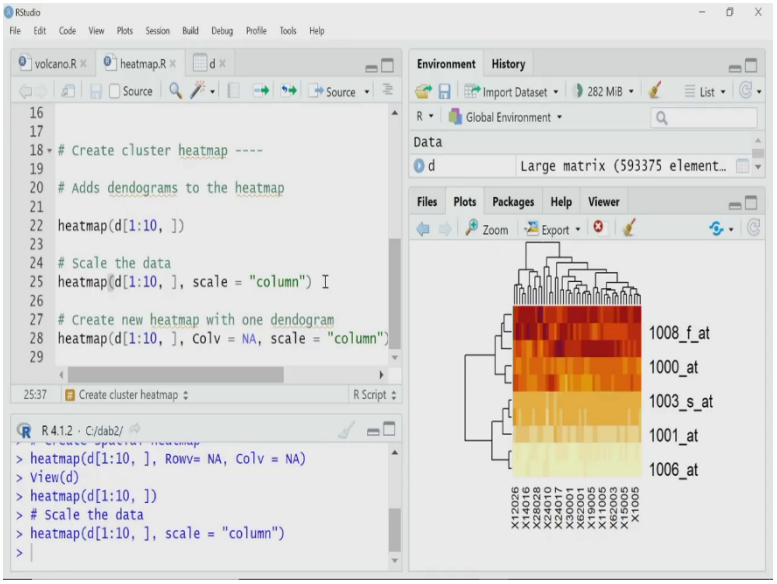
They belongs to a same cluster. Whereas the cluster created by these two genes, also belongs to another bigger cluster, which has 1005 underscore at. So in this way, it builds a hierarchical cluster and show that clustering information by your dendogram. We will have a separate lecture where I will discuss about the basic mathematical principles of hierarchical clustering and I will also have a lecture where I will show you how to draw the draw this dendogram.

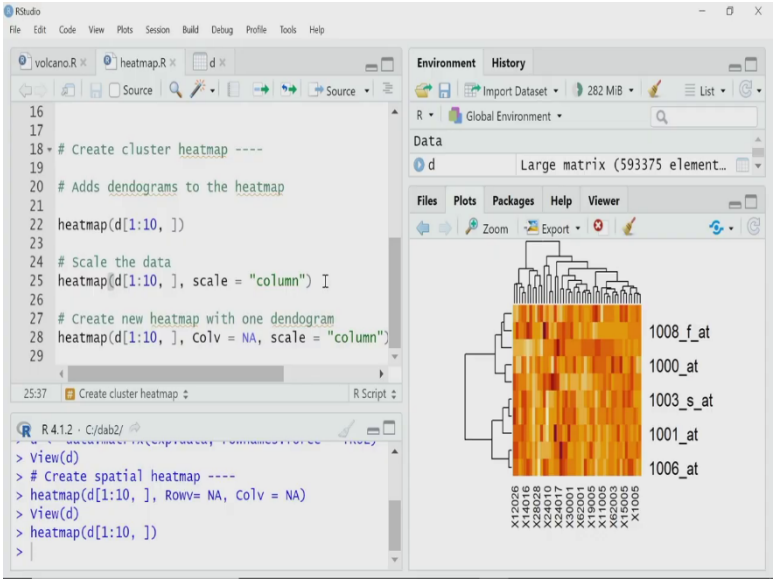
For the time being, let us use the way heat map has provided function, has provided as a dendogram. Heat map function has some arguments by which you can manipulate those properties of that dendogram and the clustering, you need not to go to them right now, whenever you will require for your particular experimental purpose you may look into them.

Now, looking into this data, I have drawn dendogram, but I do not clearly really see any cluster. Cluster means there should be gene expression of these clusters will have similar behaviour or something like that.

(Refer Slide Time: 11:54)







heatmap(d[1:10,], scale = “column”)

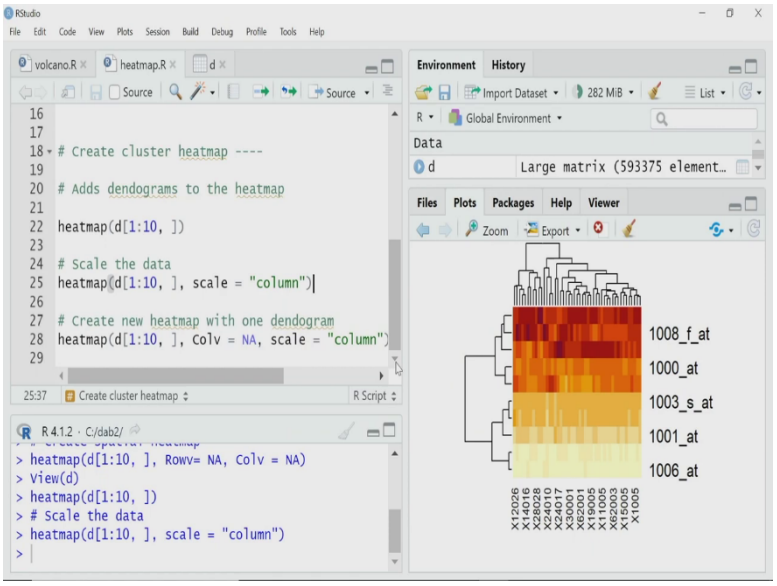
We may not be getting that cluster information, because we have not scaled the data. By scaling I mean, that I if suppose, I say a scale along the column, column is here in this heat map is one particular sample. And the rows are the genes or the features of the microarray. If I say scale the data for each column that means that I will take the data of a column that means it is coming from one sample.

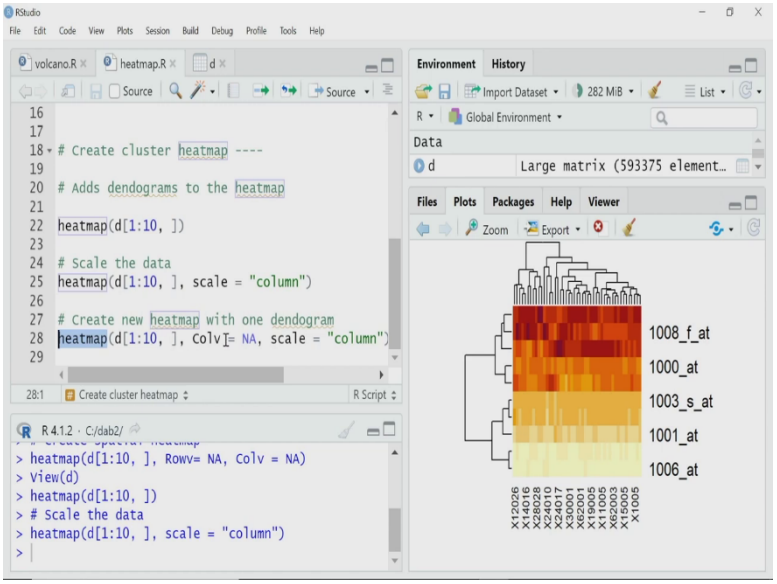
And then I will scale each of the features data in that column in such a way that the mean will become zero. So, I will centre the data. So, what I am doing here in the next line, line 25, is that I am calling that heat map function again. And I want to draw the diagram with dendogram. But now I want to scale the data for each column.

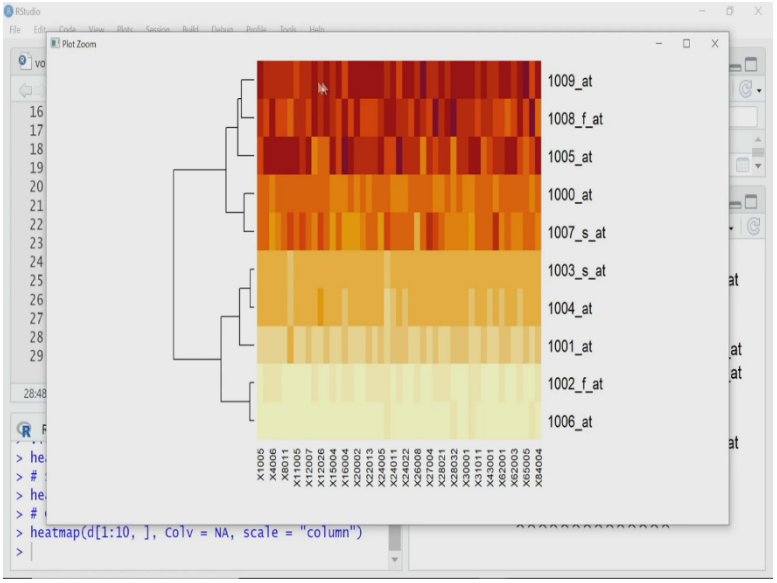
So, I am written scale equal to column, if I execute it, you can see there will be changed in the diagram. Now, the heat map has a clear colour representation of each of these clusters. You can see for 1009 underscore at, and 1008 underscore t, the readings are very high for all the samples. And that is why they belong to the same cluster, as you can see by this dendogram here at the end.

Then similarly, if you go down 1002 underscore f, underscore at, and 1006 underscore at, these two features, they have very low reading once you have scaled the data across all the samples, and that is why they belong to an individual cluster. So, now, as I have normalized or scaled the data, the clustering for these gene becomes much more meaningful. I could have done the scaling across these rows also, but in that case, in particular for this experiment, this way of scaling is much more meaningful.

(Refer Slide Time: 13:56)







heatmap(d[1:10,], Colv = NA, scale = “column”)

Now, once I have scaled and plotted this one, you may wonder this actually, I do not want to know the clustering information for the samples, the samples are clustered in the upper histogram here, which is very dense.

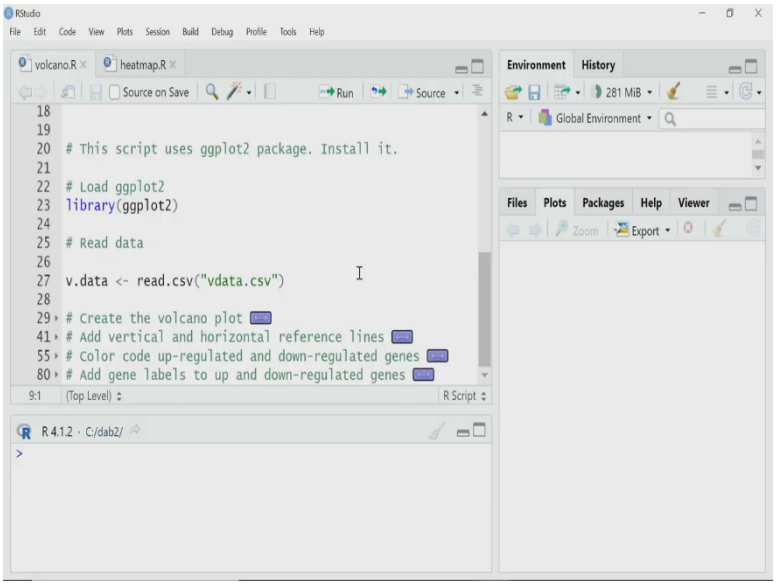
So, I want to remove that. So, what I am doing here I am saying, I am again drawing the heatmap using the heatmap function, and I am saying colv equal to NA. So, the columns which are represented in the upper dendogram will not be there right. So, that dendogram will be removed. So, I will remove that dendogram.

I execute that now the picture is much more cleaner, I have 10 genes or 10 features in each row and each column which are named below in the horizontal axis are samples. And on the

left hand side, I have the dendogram for all these genes and the colour coding is scaled colour coding and in this case the colour coding the lower readings are represented by faint colour, whereas the higher ratings are represented by the dark colour.

Heat map has options by which you can change this colour coding also, but for this video I am not going to details of that. That is all for heatmap. Now, I will move into volcano plot. Another very useful tool for representation, overall representation of high throughput data, particularly data coming from gene expression experiment like microarray or RNA seq data.

(Refer Slide Time: 15:31)

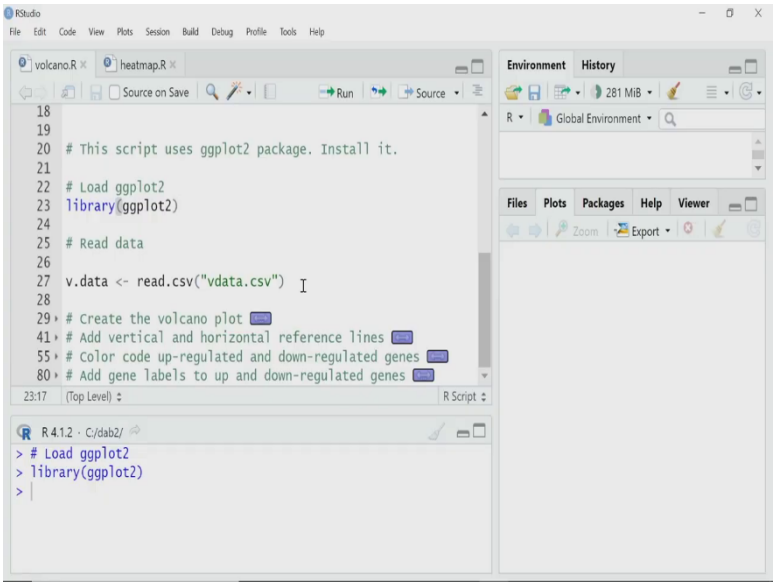


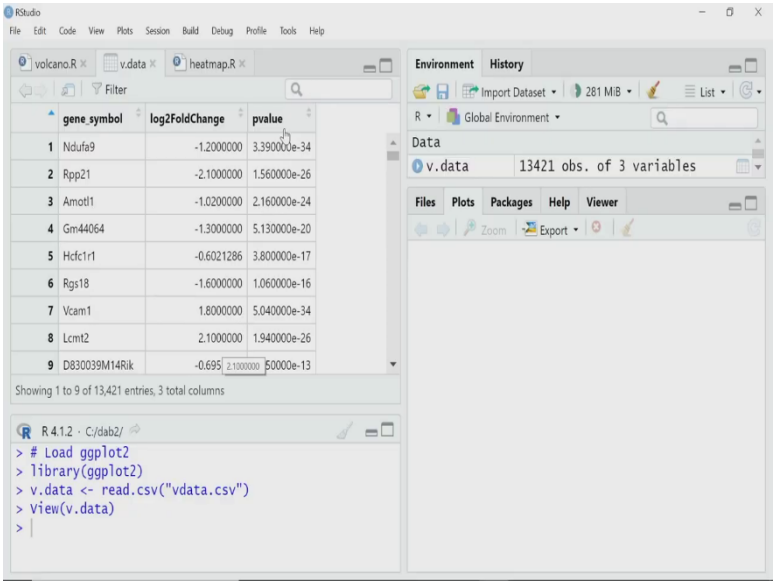
library(ggplot2)

v.data ← read.csv(“vdata.csv”)

To draw the volcano plot, I will use gg plot 2, a special package for graphical representation of data, which is very powerful. We will discuss some feature of ggplot2 separately in another video, but here, you just follow me how to use ggplot2 function for particularly drawing the Volcano plot. So, you need to install ggplot2. Hope you may have already installed it. And then you have to call ggplot2 from your library to unload it in your workspace before you execute it.

(Refer Slide Time: 16:07)





library(ggplot2)

v.data ← read.csv(“vdata.csv”)

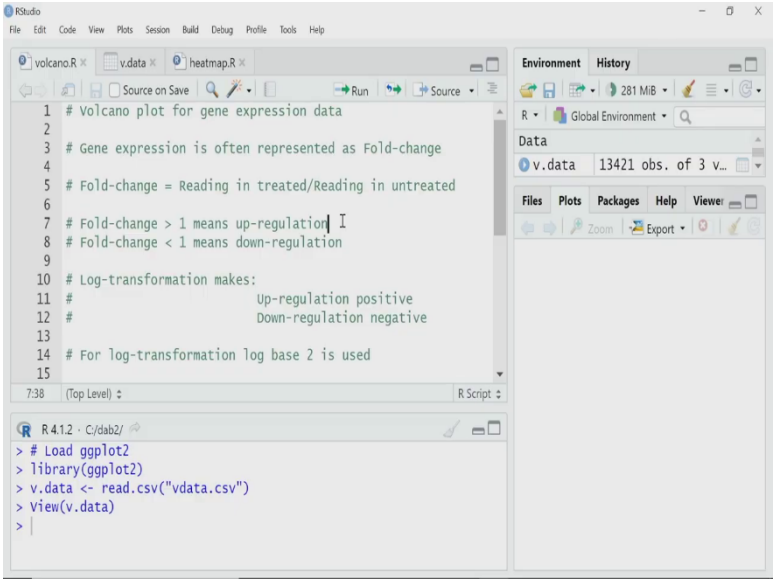
So, the first line of my script is doing that. Now I will read the data. And then I will explain what type of data I have and why I want a volcano plot and what is the volcano plot. So, the data is in a CSV file, and I will read that let me open this data to see what I have and what I want to plot. The first column of this data is gene symbol, this can be a microarray feature name also. The second column is log 2 fold change.

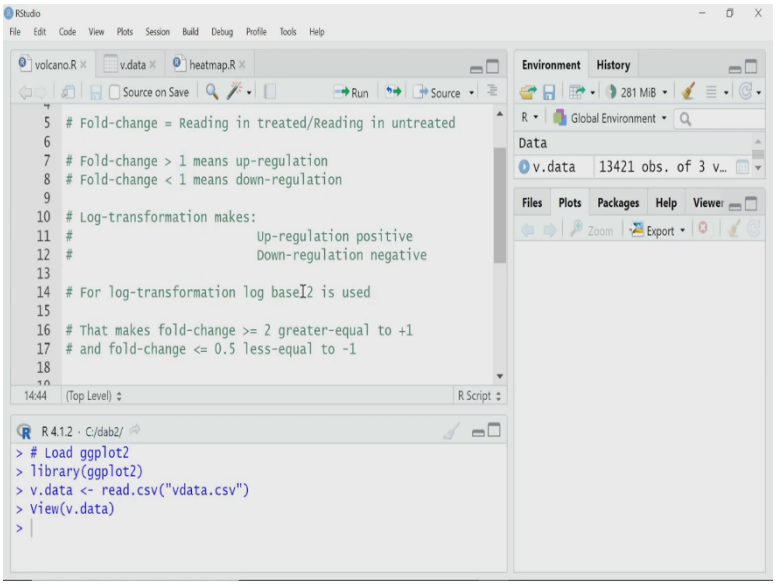
So, that is fold change in, transformed in log two, log base two. And the third column, the last column is p value. Now, let me first explain what is log 2 fold change. Suppose you are doing

a microarray experiment or even if you are doing a small scale quantity PCR experiment or a large scale RNA seq experiment, you always have a reference sample.

Many a times that can be untreated sample and your experimental sample can be the treated sample whereas in some cases you may have a normal cell, whereas that will be the reference sample whereas, the cancer cell or tumor will be the experimental sample. Now, when you measure gene expression, either by quantitative PCR or microarray or RNA seq, you always represent the data in terms of fold change, relative measure with respect to which is in my reference sample, that is untreated sample.

(Refer Slide Time: 17:48)





So, I can say the fold change as I have written here is equal to reading in treated divided by reading in untreated reading in the disease cell divided by reading in the untreated cells. Now, you can easily understand fold change can be bigger than one; fold change can be less than one. When fold change is bigger than one that means a reading in treated sample the expression of the genes in the treated sample is more than the expression in the untreated sample. So, that is called upregulation of that gene in that treated sample.

Whereas, if one, if the fold change is less than one you will call it downregulation it is easy to remember but it is not very intuitive. So, what usually we do in high throughput data analysis is that we log transform that fold change data, that means, I take a log of that, then what will happen if I take a log of that any raw value, any raw fold change value if it is bigger than one. Then it will remain a positive value. Whereas, if you take a fold change value which is less than one less than one mean it is fraction right. So, if I take a fraction and take a log of that, I get a negative value. So, if I log transform the data and then if I sort the data in Excel sheet or in our anywhere, you can easily recognize, the positive values are upregulated gene.

The negative values are downregulated gene. It is much more intuitive for us to understand. Now, when you are taking log, you can take any base you can take 2, 10, 100 whatever value you want as a base. Now, for microarray and RNA seq this type of experiment where you are doing gene expression data. Many times conventionally we take 2 as a base, why do you do that? It is a convention.

See, I may have done an experiment suppose by quantitative PCR also simply by real time PCR machine in our lab, you have seen the expression of a particular gene has changed 1.2 fold and you have done statistical tests and that change is consistent and statistically significant. But remember, one point two fold change in gene expression may not affect the biology of the process.

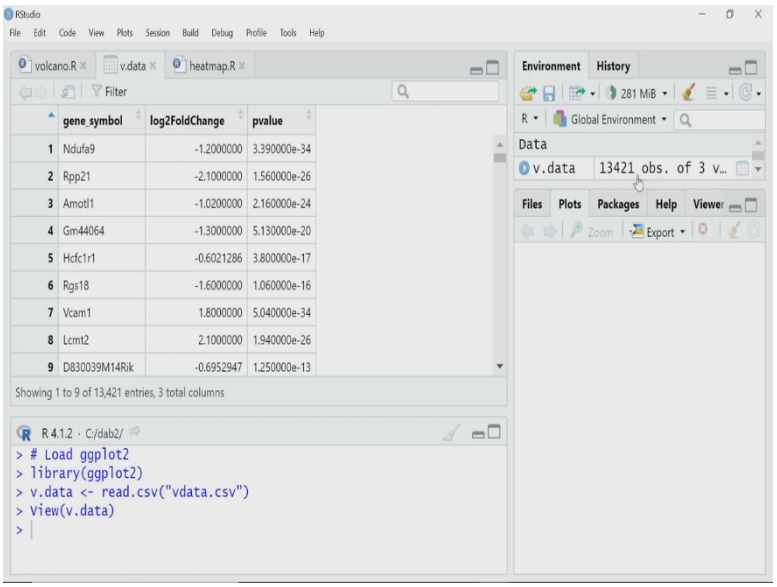
That means, although this data is statistically significant, and you repeatedly actually get some something close to 1.2 or 1.5, but that does not mean that fold change has any relevance in the biology that you are studying. So, many times within the community, biology community, we usually consider that any change any fold change, it is twofold, either two fold upregulation or two fold down regulation then only there is any biology relevance.

That means a fold change has to be bigger than two fold, so if it is upregulated, then it should be 2 or more than 2. If it is downregulated, it should be 0.5 or less than 0.5. So, that is why

when we do the log transformation, we take log base two, then what will happen if the fold changes originally is 2 log 2 of that will give me plus 1.

And if the fold change is point five, it will give me minus 1 and easily you can easily identify which one is upregulated and downregulated from the sign and from the value, I can understand whether those upregulation and downregulation are biologically relevant or not.

(Refer Slide Time: 21:25)



Now, let me go back to my data again. So, this first column is for Gene symbols second column is log 2 fold change of for these gene, each of these gene and then third column, the last column is for P value, what is this p value? They must have, I have taken data from a data set, they must have perform some sort of statistical test to check whether this change in gene expression between the treated and untreated disease and the normal is actually statistically significant or not.

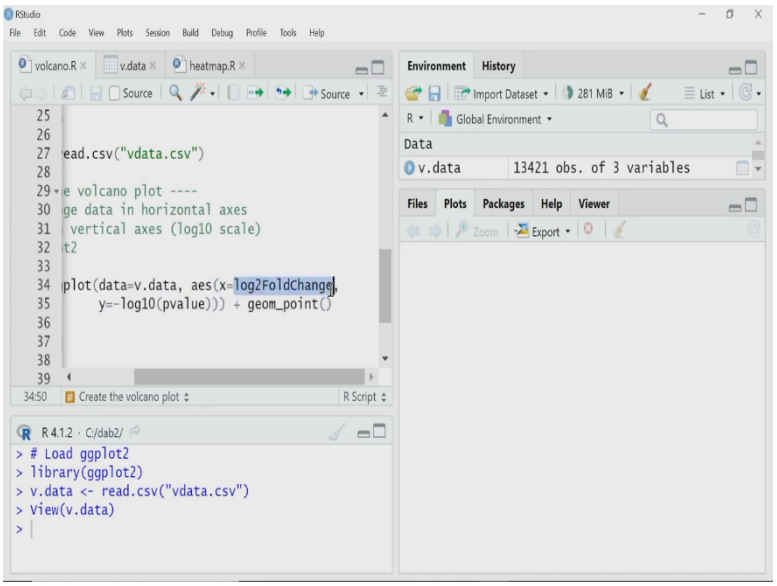
And from that test, they have calculated some p value. If the p value is small and below a cut off, I will say it is statistically significant. So, that raw p value is given in the last column. So, now I want to create a plot where I will represent these fold change in log 2 scale and the p value for each of these gene in one single plot.

Remember, in this example, for example, I have 13,421 observation, that means I have data for 13,421 genes for each gene, I have log 2 transform fold change data and the corresponding p value data. And I want to visualize that in one single diagram in such a way

that I can understand which are the genes or how many of these gene are actually upregulated and downregulated.

And such upregulation and downregulation is biologically relevant and statistically significant. Two things together, I want to visualize biological relevance, as well as statistical significance for all of these gene expression data change in one single plot. For that, I will use volcano plot.

(Refer Slide Time: 23:15)



plot1 ← ggplot(data = v.data, aes(x = log2FoldChange,

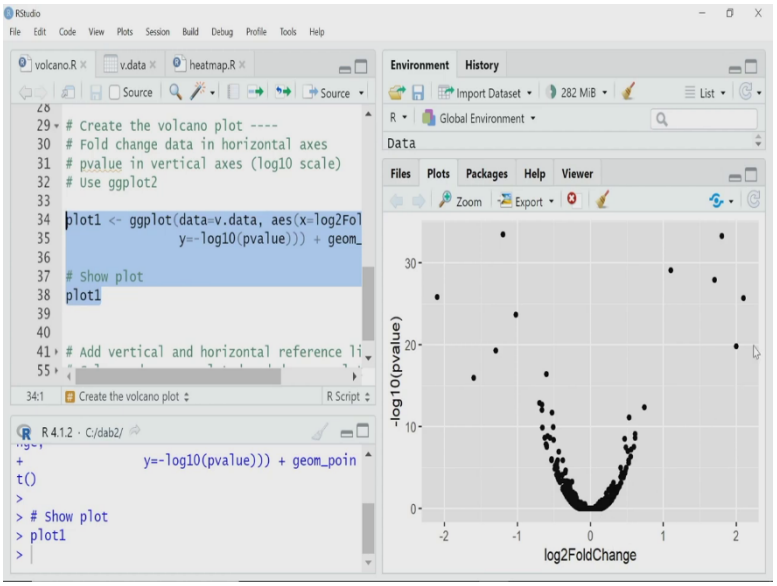
y = - log10(pvalue))) + geom_point()

I will be using gg plot to create the volcano plot. Let me explain briefly what are the commands that I am using to do that. So, the first thing is I am calling gg plot function. And then as an argument I am telling the data is v dot data. And then I have something written as aes, it stands for aesthetic. We will discuss in detail about these when we will discuss the gg plot separately.

So, aes function, I am saying that, in the x axis, the horizontal axis, you put the log 2 fold change data. And I am also saying in Y axis, you put minus log 10 of P value. So, I am not plotting in vertical axis the p value directly, I am taking a log 10 transformation of that, that means log of p value base 10. And I am putting a minus sign before that, because if you remember, these p values are fractions.

So, if I take log it will become negative. So, that to make that value positive, I am making it minus. So, the bigger this value is, I have higher statistical significance. So these two will be the axis and then I am writing plus, I am adding another property to this plot, where I am saying geom underscore point. That means gg plot will understand that I want to create a scatter plot. So, that information will be stored in plot1 object and then I will plot that

(Refer Slide Time: 24:50)




```
plot1 ← ggplot(data = v.data, aes(x = log2FoldChange,
                    y = - log10(pvalue))) + geom_point()
plot1
```

So, here is my volcano plot. It is called volcano because it looks like eruption from a volcano, where the mouth of the volcano is here, near zero in the horizontal axis. Each of these black dots is a scatterplot is those, one of those 13,000 genes and they are positioned based upon their log 2 fold change value, which is coming from the second column of my data. And the vertical axis is minus log 10 P value.

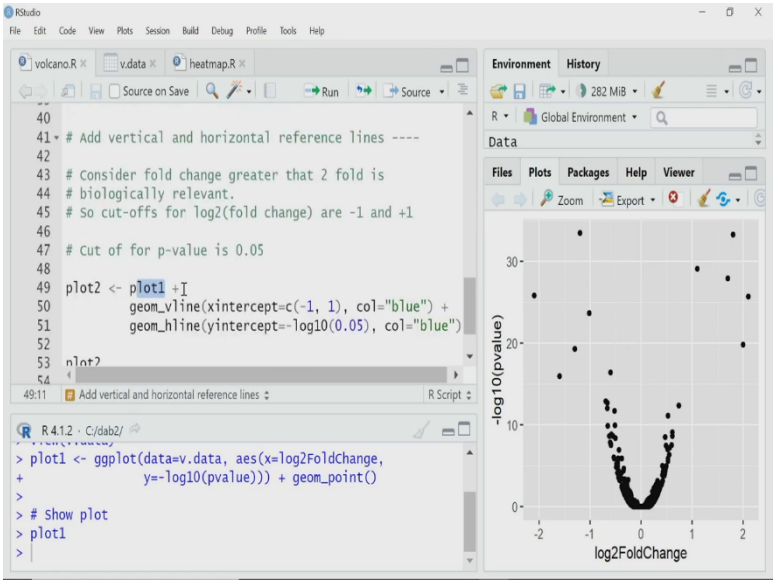
For example, this higher one this has a p value which is very small. That is why the minus log 10 value is very high. And its fold change value in log 2 scale is somewhere near minus one slightly smaller than minus one. So, now, this is a volcano plot. In these all these 13,000 genes has been shown.

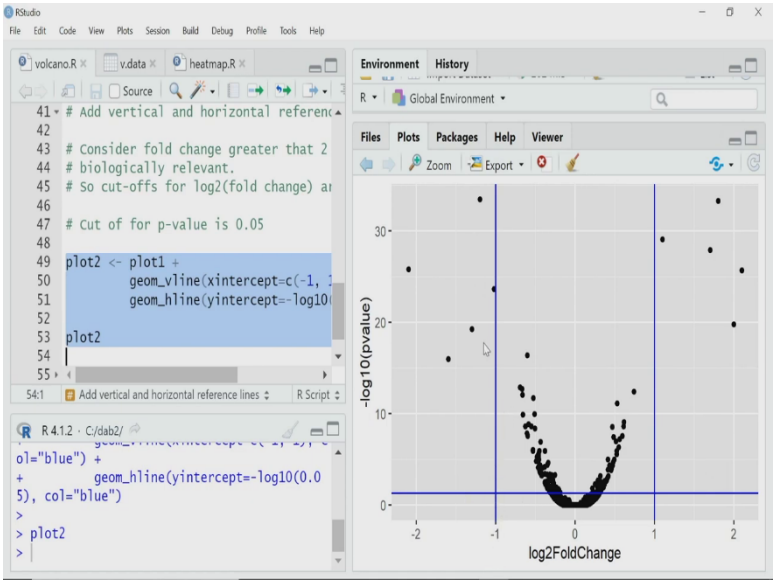
And there are two properties, ones, whether they are statistically significant or not, the under fold change is biologically relevant or not, I want to visualize. How do I want to visualize? Any gene whose expression is bigger than one that means log 2 fold change is bigger than one on the right hand side, I should consider that they have a biologically relevant fold change.

Whereas any genes having log 2 fold change lesser than or equal to minus one, I will say they have biologically relevant downregulation. Similarly, along the vertical axis, I have to check, draw line which correspond to suppose, I take a cut off of 0.05. So, minus log of 0.05 base 10 that should be my cut off in the vertical axis and any data above that represent statistically significant data.

So, let me draw these vertical and horizontal lines; one vertical line to represent the cut off for the p value, whereas one horizontal line for the cut off for P value, and two vertical lines to represent the fold change upregulated or downregulated.

(Refer Slide Time: 27:10)






```
plot2 ← plot1 +
        geom_vline(xintercept = c(-1, 1), col = "blue") +
        geom_hline(yintercept = - log10(0.05), col = "blue")
plot2
```

To achieve that, what I will do, I will modify this plot by adding some vertical and horizontal line. So, my cut off for the, in the horizontal axis is minus 1 and plus 1, whereas in the vertical axis, the cutoff is 0.05. So, I am taking the plot1 if you remember we have created the plot1 object here, which is shown here in the plot tab. Now, I will take that plot1 and on that I will add two vertical line and one horizontal line.

And in gg plot it is very easy to do by simply add a symbol I will add the vertical line and the horizontal line, what will be the vertical line I am calling the function geom underscore V line and then I am saying draw a vertical line having intersect at minus 1 and plus 1 and the colour of that vertical line should be blue.

Not just that, I am putting another plus sign to represent that, I want to put another argument here, I want to draw a horizontal line for that I am calling geom underscore H line and then I am saying horizontal line the y intercept for that should be minus log 10 of 0.05. So, my cutoff of p value is 0.05. And I want the colour of the line is also blue.

And I will plot that. This is my new plot. Now it is much more meaningful. I have two vertical line one is at plus 1 one is at minus 1, any value starting from plus 1 and higher are genes which are upregulated. So, I can see here I have five genes which are upregulated. And

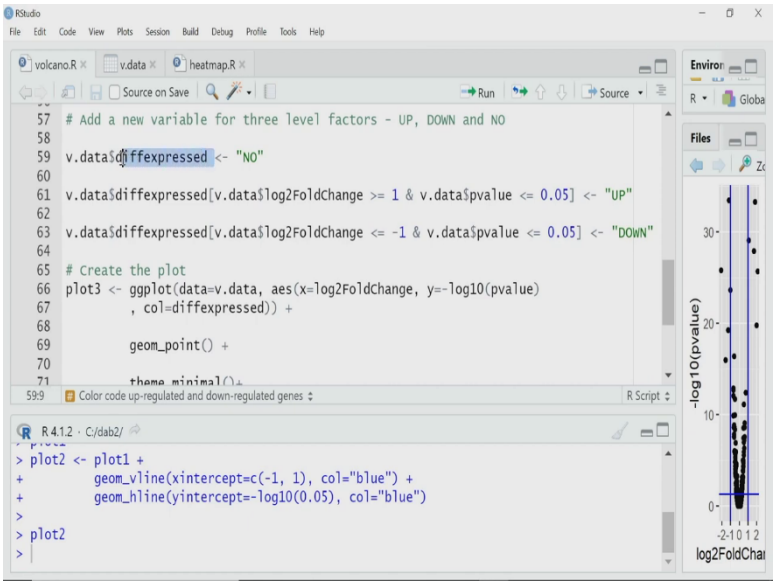
I have another vertical line at minus 1 any genes at minus 1 or below minus 1, on the left hand side of this vertical line, are biologically relevant downregulated.

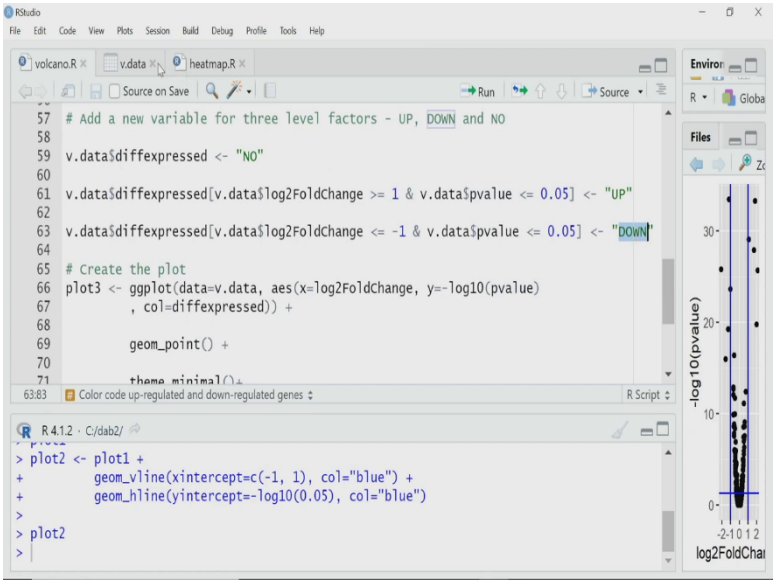
Whereas the cut off for P value 0.05 is this horizontal line, any data, any gene fold change data above that line have a statistically significant change in fold of expression. So, if I look at it comprehensibly, on the right hand side, I have 5 genes which have biologically relevant upregulation as well as their change in gene expression are statistically significant. Whereas on the left hand side again I have 5 genes which have biologically relevant downregulation, at the same time they are statistically significant.

Rest of the genes, rest of those 13,000 something genes, many of them has statistical significance. That is why they are above this horizontal line of cut off at point p equal to 0.05. But they are not biologically relevant, their change in fold expression, fold change are not biologically relevant. So, I will discard them from my subsequent analysis.

I will focus on these 5 and 5, 10 genes. Once I have these Volcano plot, now, I may want to add some further element on it. For example, I may want to color code the upregulated genes and the downregulated genes. How can I do that, I will show that.

(Refer Slide Time: 30:50)





v.data$diffexpressed ← "NO"

v.data$diffexpressed[v.data$log2FoldChange >= 1 & v.data$pvalue <= 0.05] ← "UP"

v.data$diffexpressed[v.data$log2FoldChange >= -1 & v.data$pvalue <= 0.05] ← "DOWN"

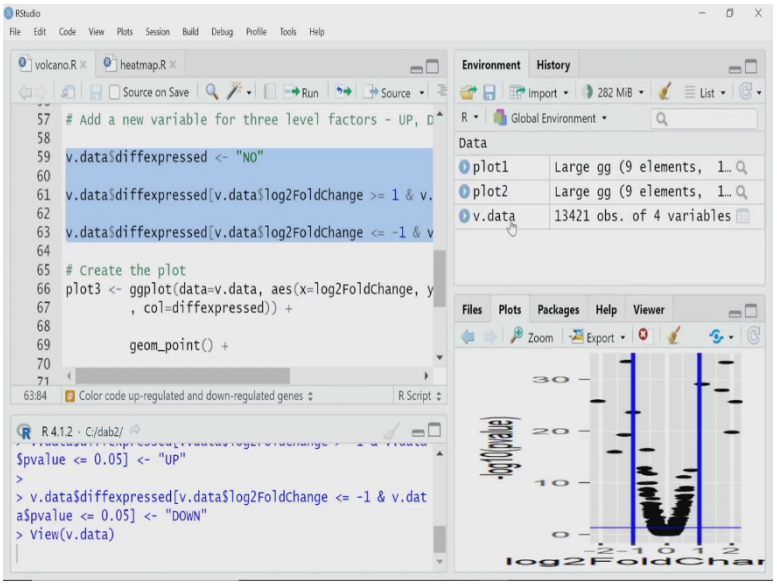
To mark those genes, which are upregulated, and downregulated, at the same time, they are statistically significant in their fold change, I had to create a new variable, a new column in my data set. So, what I am doing here, I have v dot data. And there I am adding a new column called diffexpressed differentially expressed, and I am assigning no to them. That means a new column will be created, where all the entries will be no at first.

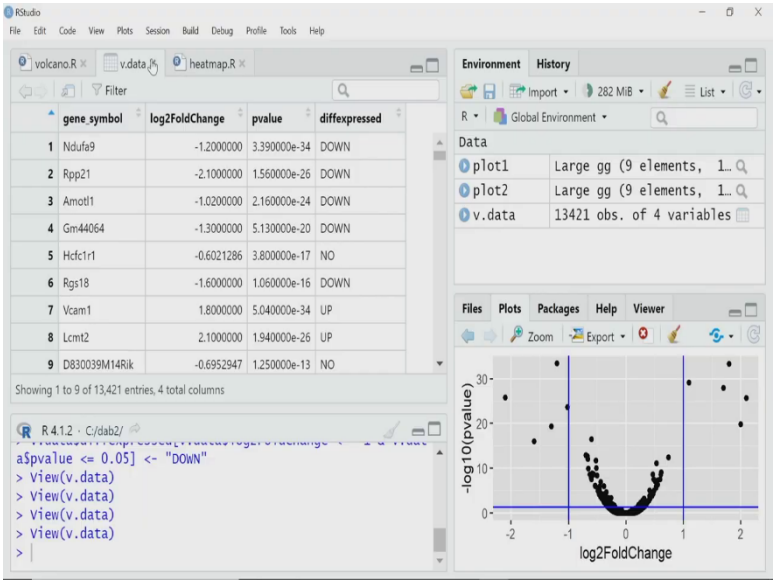
Now, in the next line, what I am doing, I am again, calling the same thing. I am saying v dot data and putting the dollar sign and I am calling this variable diffexpressed, which I just created and put no values there. And I am taking that column, the data of that column, and then what I am doing, I am asking to check some logic. What is the logic?

If the data has those rows where log2Fold change is greater equal to 1, and where the p value is less than 0.05. So, that means now R will go to all those rows of my data, where log2Fold change is greater equal to 1, and the p value is less equal to 0.05. Then, in this last column that I created diffexpressed for those rows in that diffexpressed, I will replace these NO by UP, whereas the next line, I am doing the same thing, but I am giving a different logic.

What I am giving a logic, my logic here is go to all those rows were log2Fold change is less equal to minus 1, and the P value is less equal to 0.05. Then in the diffexpressed column for those rows you write DOWN.

(Refer Slide Time: 32:50)






```
v.data$diffexpressed ← "NO"

v.data$diffexpressed[v.data$log2FoldChange >= 1 & v.data$pvalue <= 0.05] ← "UP"

v.data$diffexpressed[v.data$log2FoldChange >= -1 & v.data$pvalue <= 0.05] ← "DOWN"

plot3 ← ggplot(data = v.data, aes(x = log2FoldChange, y = - log10(pvalue)
        , col = diffexpressed)) +
```

```
geom_point() +
theme_minimal() +
scale_color_manual(values = c("red", "black", "green")) +
geom_vline(xintercept = c(-1, 1), col = "blue") +
geom_hline(yintercept = - log10(0.05), col = "blue")
```

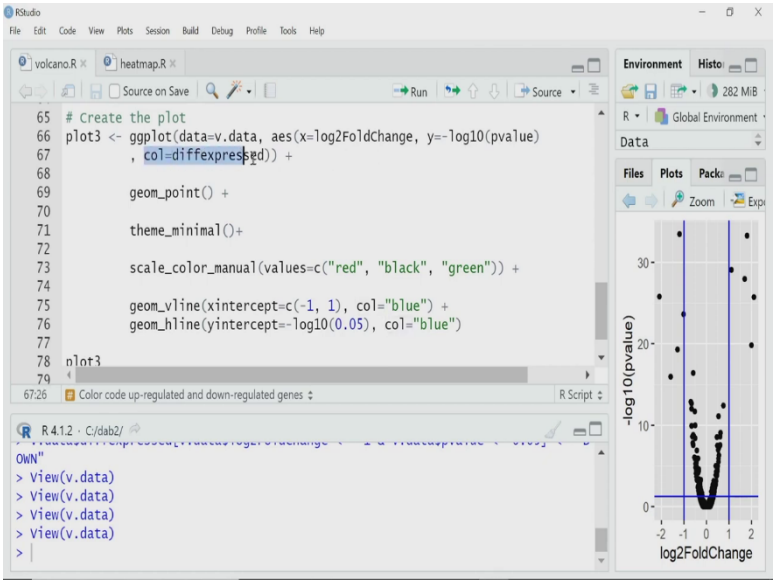
plot3

So, if I do that, let me execute that. I have executed it. Now I will go back and check my data file, v dot data. Now you can see I have a new column diffexpressed and in this column, some genes are written as DOWN; see the Fold change value is minus 1.2. So, it is lesser than minus 1, and the p value is much smaller than 0.05.

So, it is statistically significant and biologically relevant down regulation that is why DOWN is written here. Whereas for the Vcam1, the Fold change is 1.8, Log 2 Fold change. So, it is bigger than plus 1 and the p value is also satisfying our requirement, so it is marked as UP. Whereas this fifth gene, we are marking as No, because this gene expression change fold change is within plus 1 and minus 1.

So, that fold change is not biologically relevant. And that is why although the statically p value is good enough, it is marked as No. So, now I have a new column where each gene is marked either as Down, Up or No. Now I will use that information to color code my volcano plot.

(Refer Slide Time: 34:25)

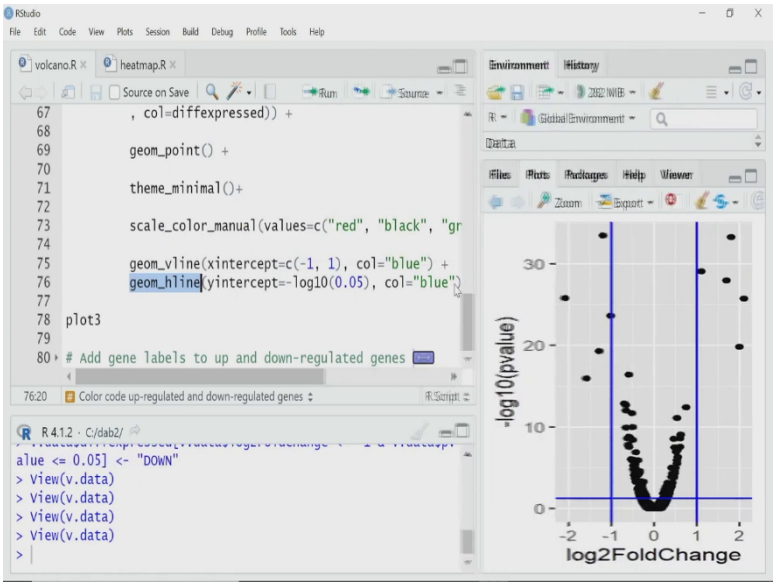



```
plot3 ← ggplot(data = v.data, aes(x = log2FoldChange, y = - log10(pvalue)
                , col = diffexpressed)) +
                geom_point() +
                theme_minimal() +
               scale_color_manual(values = c(“red”, “black”, “green”)) +
                geom_vline(xintercept = c(-1, 1), col = “blue”) +
                geom_hline(yintercept = - log10(0.05), col = “blue”)
```

```
plot3
```

To create the volcano plot, I have added lots of feature along with the ggplot basic option that I have used earlier. So, the first thing is ggplot I am using I am saying the data and I am using the same aesthetic, where x equal to Log2Fold change and y is minus Log10 of pvalue. Now I have added another argument here in aesthetic that is color, color equal to diffexpressed. That means I am asking that you color these data points based on the values in this column called diffexpressed.

(Refer Slide Time: 35:04)



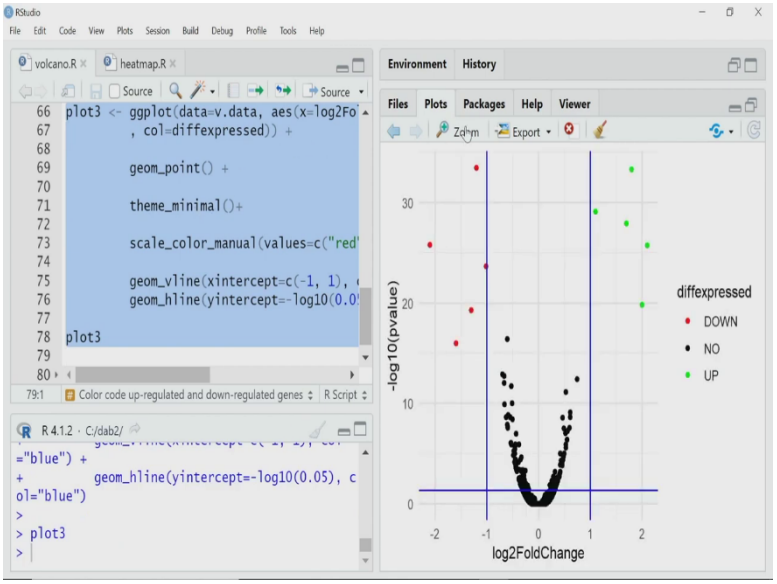

```
plot3 ← ggplot(data = v.data, aes(x = log2FoldChange, y = - log10(pvalue)
                , col = diffexpressed)) +
                geom_point() +
                theme_minimal() +
                scale_color_manual(values = c("red", "black", "green")) +
                geom_vline(xintercept = c(-1, 1), col = "blue") +
                geom_hline(yintercept = - log10(0.05), col = "blue")
```

```
plot3
```

And then I am asking to create the scatterplot just like the previous one. Here, I have made a change, there are multiple themes in ggplot 2. So, I am saying use the minimal theme. So, these background thing will get bit cleared. You can skip that. Then I am saying, you have to manually color that do not use a default color, what should be the color, the value should be red, black and green.

Then obviously, I am drawing the vertical lines for the gene expression Fold change cut off and the horizontal line for my p value cut off. So, let me execute it and draw the plot.

(Refer Slide Time: 35:44)



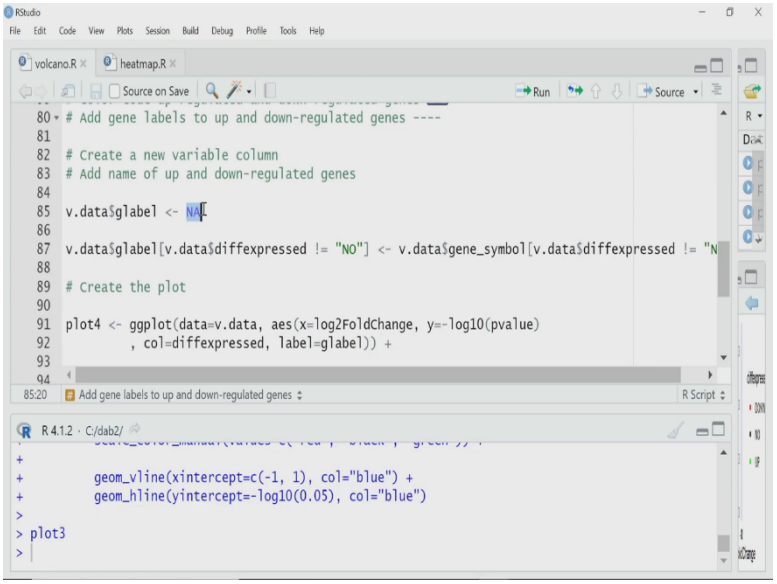

```
plot3 ← ggplot(data = v.data, aes(x = log2FoldChange, y = - log10(pvalue)
        , col = diffexpressed)) +
        geom_point() +
        theme_minimal() +
        scale_color_manual(values = c("red", "black", "green")) +
        geom_vline(xintercept = c(-1, 1), col = "blue") +
        geom_hline(yintercept = - log10(0.05), col = "blue")
```

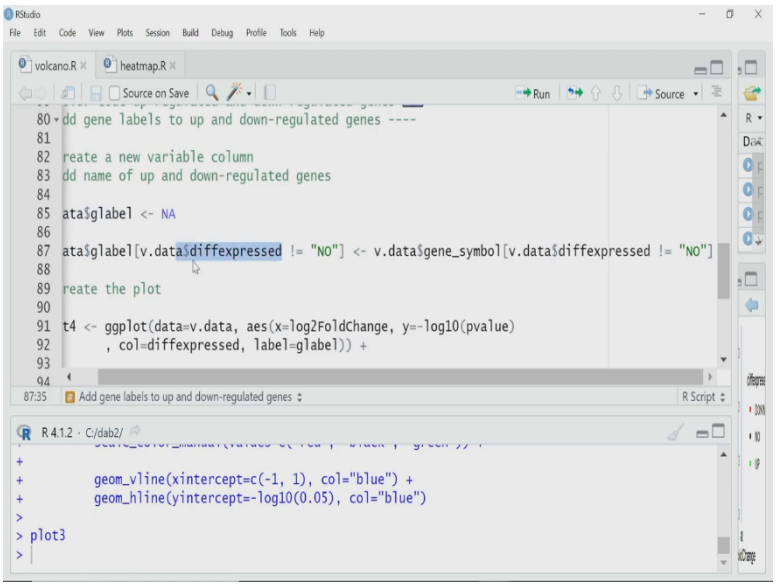
plot3

Now, you can see the picture, the plot work on the plot is much more clear. So, I have all the data points, all these 13,000 data points. And I have shown the cut off by these horizontal and vertical blue line upregulated genes are green colored, and the down regulated gene are red colored. And the legend has been written here so that one can understand which one is DOWN regulated which one is UP regulated.

The black color is for those gene, which does not satisfy our requirement for biological relevance and statistical significance. Now, in many diagram, many a time in many paper, you will see people will add the name or the label for each of these genes. So, now let me show you how to add the name of those gene or the label for each of these data points.

(Refer Slide Time: 36:57)





v.data$glabel← "NA"

v.data$glabel[v.data$diffexpressed != “NO” ← v.data$gene_symbol[v.data$diffexpressed != “NO”]

To achieve the labeling of those UP regulated and DOWN regulated gene, I have to play some trick with my data file, what I will do, I will add another column to my data where the label will be written. How I will do that? I am taking my v dot data. And then I am adding a new column called glabel in new variable.

So, I am writing v dot data, dollar sign glabel, and then I am assigning NA value to this, nothing NA to that. And then I will replace these NA with certain things. What I will do?

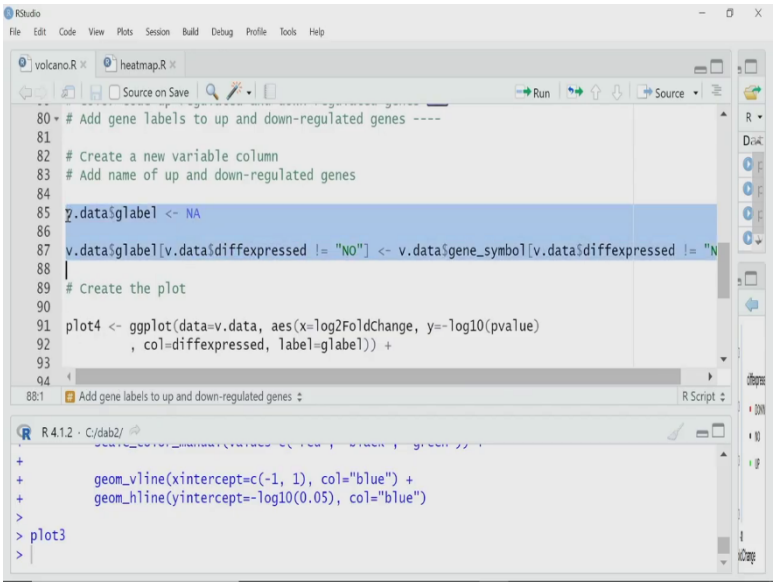
Now, once I have created that new column and filled them with NA, I am taking that column, so I am taking v dot data dollar sign glabel, so that column is taken.

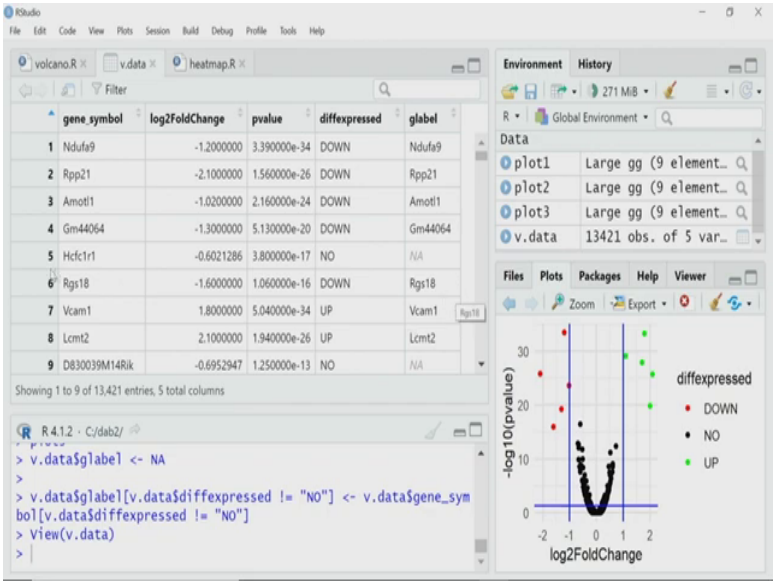
And then I am using some logic operation here. What is the logic operation? I will go to each row by using the statement here v data dollar diffexpressed, and I will check if diffexpressed is not equal to NO. Remember in the diffexpressed column, I have either written each row is marked as DOWN UP or NO.

NO means there is no change in expression, we are not interested in them, we are interested to label only those genes which are UP regulated or DOWN regulated. So, I am using the logic that find those rows where diffexpressed is not equal to NO, that means they are either UP or DOWN. And then what you do you replace this NA in those row for the glabel column with the gene symbol corresponding to those rows.

How do I get the corresponding to those rows. So, again, I use the same logic that I have used here. So, I am creating a new column where the label will be written. What will be the label? Label will be the gene names, but those who will be present only on those rows, where the differential expression is either UP or DOWN. Let me do it, it will be much clearer if I open the file and check it.

(Refer Slide Time: 39:07)





v.data$glabel← "NA"

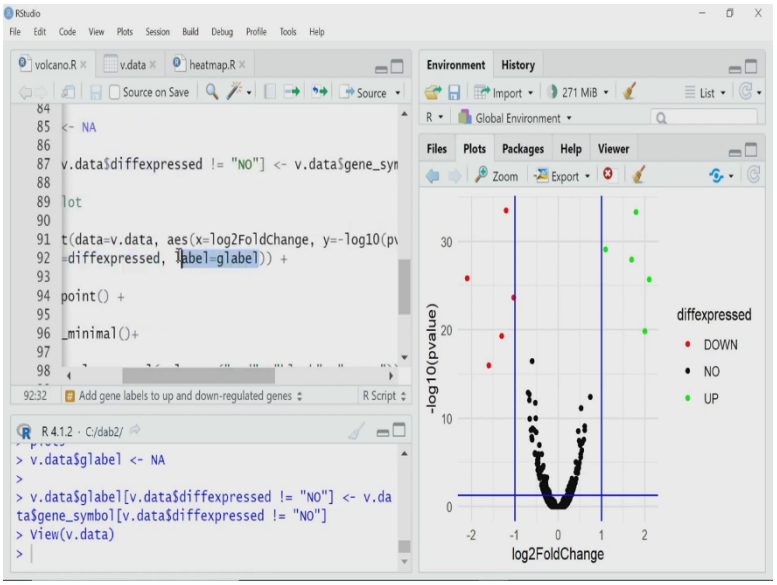
v.data$glabel[v.data$diffexpressed != “NO” ← v.data$gene_symbol[v.data$diffexpressed != “NO”]

It is done let me open the data to see what changes I have made. Now, you see, I have a new column gene label and take the fifth one. The fifth one, the diffexpressed column says NO, because it is gene expression is within plus 1 and minus 1. So, it is not biologically relevant. So, as it is know that glabel is NA.

But whereas for the other 4 genes which are above their corresponding gene name is present in the glabel for each of these row. Similarly come to the UP regulated one Vcam1 and Lcmt2. You can see the glabel row column has the corresponding gene names. So, by using that logic operation I have created a new column where the gene names for each of these UP regulated and DOWN regulated genes are listed there.

Now, I will use ggplot and ask it to use this glabel column information to write down the name of these genes on those dots, dot correspond to each of these gene in my volcano plot.

(Refer Slide Time: 40:27)



v.data$glabel← "NA"

v.data$glabel[v.data$diffexpressed != “NO” ← v.data$gene_symbol[v.data$diffexpressed != “NO”]

```
plot4 ← ggplot(data = v.data, aes(x = log2FoldChange, y = - log10(pvalue)
                , col = diffexpressed, label = glabel)) +
                geom_point() +
                theme_minimal() +
                scale_color_manual(values = c(“red”, “black”, “green”)) +
                geom_vline(xintercept = c(-1, 1), col = “blue”) +
                geom_hline(yintercept = - log10(0.05), col = “blue”) +
```

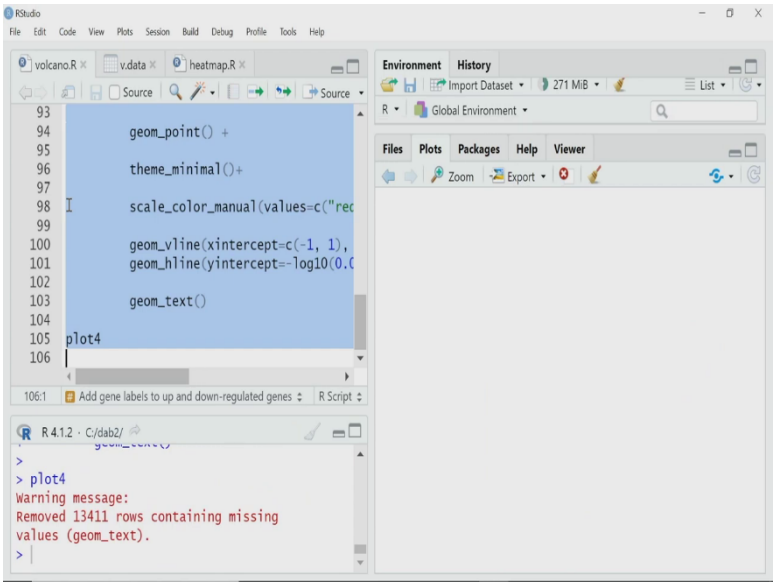
geom_text()

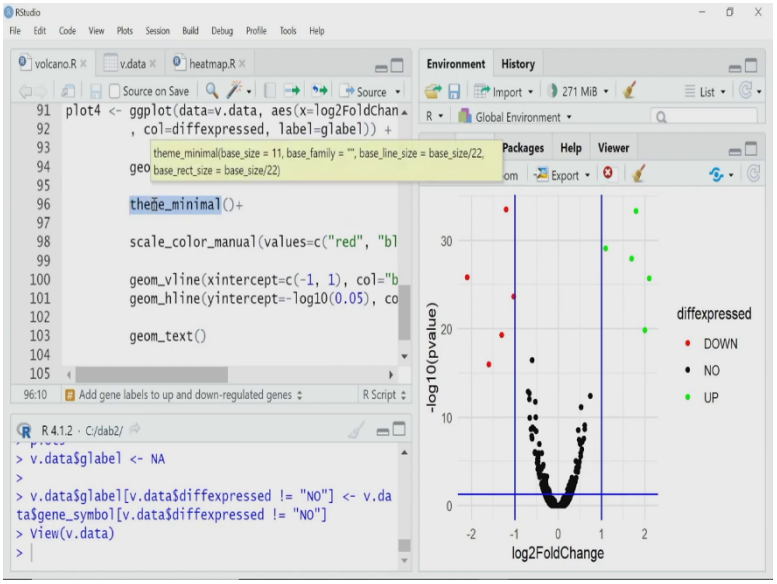
plot4

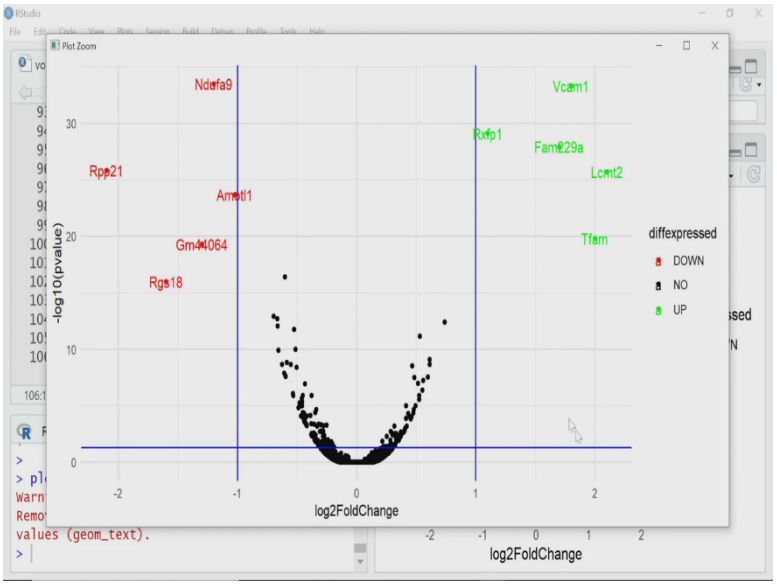
So, now, I will create the plot I am again using ggplot, and you have the usual all statements for argument like the data, the aesthetic, and in the aesthetic in the previous example, I have used color by differentially expressed information. Now, I have added a new argument in aesthetics, it is a label equal to glabel.

That means, gg plot will understand that, I have to label the data points based on the information present in the glabel column of this data set. Now, I have to make another change, I will come to that. So, the rest of the thing like geom point, that means I have a scatter plot, I am maintaining the theme minimal, so the background color is off, the color scale again, scale of the color is manual. So, I am saving red, blue and green, the way I have just done now. And then I have the vertical lines and the horizontal line.

(Refer Slide Time: 41:35)








```
plot4 ← ggplot(data = v.data, aes(x = log2FoldChange, y = - log10(pvalue)
                , col = diffexpressed, label = glabel)) +
                geom_point() +
                theme_minimal() +
                scale_color_manual(values = c(“red”, “black”, “green”)) +
                geom_vline(xintercept = c(-1, 1), col = “blue”) +
                geom_hline(yintercept = - log10(0.05), col = “blue”) +
                geom_text()

plot4
```

The last one is a new one. This is geom underscore text function, this will tell the ggplot that you have to write the labels, labels based on the information present in glabel column. Remember, if it is NA ggplot, we will not write anything for that particular data. So, I will execute it. I have the volcano plot now, which looks quite professional.

So, this is a neat volcano plot where all the 13,000 genes data are presented, each of these gene is a dot, they are color coded. If they are black, that means they do not satisfy the biological relevance and statistical significance. That is why they are marked as No. Whereas the green dots for which I have the name written here, Vcam1 Fam229a Tfam like that.

These are the UP regulated genes. And they satisfy both the criteria that the Fold change should be bigger than 2, as well as their Fold change should be statistically significant. And that is why they are above the horizontal line and beyond this vertical line for one, whereas these red genes for which we have written the name, for example, Rpp21 is one example.

It is on the left hand side of the minus 1 vertical line. So, that means its Fold change is biologically relevant. It has a DOWN regulation as well as its DOWN regulation is statistically significant, because it is above this horizontal line of cut off p equal to 0.05. So, now in my subsequent study, I can focus on these 5 and 5, 10 genes and build new experiment or perform some new analysis.

So, that is all for this lecture on volcano plot and heat map. There are many dedicated tools to draw heat map and volcano plot which comes with packages which has been created for analysis of microarray or RNA Seq data. If you are using those packages for detailed analysis of your data, microarray RNA Seq data starting from the raw data up to different diagrams, then you may use those tools.

There are inbuilt tools to draw the volcano plot and the heat map. Some of them will have additional utilities and additional features. Here in this lecture, I have shown you the inbuilt heat map function and the scatter plot option available the ggplot 2. That is all for this video. Thank you for learning with me today.