Final Project

Applied Biostatistics for Bioinformatics

Devin McAvoy (dm5153), Mudra Patel (mp6092), Jessica Schilter (js12845), Brandon Thong (bt1194)

## Full code available here: <https://github.com/Celsius452/BiostatsFINAL>

# Data set 1.

(Brandon and Jessica)

**Prompt:**

*This data is about basketball players from the year 2008 and is in the file ppg2008.csv. It has various statistics on players in the NBA. You might not know what each of the metrics means (I don't), but they are just different dimensions of data. This is a visualization and data mining exercise.*

*What can you say about this dataset, use tools that you learned here. and make a report or a visual that highlights something interesting, maybe compare players, especially how they have performed since, based on the data in here. Many of these players have reached their peak recently and you will be able to find statistics about their performance in 2019.*

*Could you have predicted the successes and failures of some of the players, based on analyses of the data? Maybe you could be a talent scout for an NBA team?*

*Think of it as your job, as a reporter for NY times, to make a single graphic that highlights something about this data. Explain the analysis that went into the graphic and present the code too. This should be done in a notebook, so it is easy to evaluate.*

**DISCUSSION:**

**2008** (note that data for both sets will be of the top 50 scoring players)

**Heatmap of 2008 Players

Description automatically generated**

Figure : Heatmap of 2008 Players

At first glance in the 2008 players, 3-point shooting seems to be much less prioritized, with the rest of the players having lower but mostly equal shooting percentages for this statistic. However, in the 2018 data, we see that while 3P% is at mostly consistent levels across players, they are grouped much higher (darker colors) compared to the older data. We can see the main excellent shooters in 2018, unsurprisingly Steph Curry, Klay Thompson, and Danilo Gallinari. HOWEVER, it should be noted that it may be difficult to glean information as effectively from the 2008 data, as Yao Ming apparently made 100% of his 3pt shots (virtually no attempts, as he was an extremely tall player, even for a center, which usually did not shoot at the time).

After further investigation, it seems the issue extends beyond just Yao Ming. It seems that many large players in the Center position have extremely high or low 3-point shooting percentages due to their very low attempt numbers. However, removing Yao as an outlier still improves visualization of 3-point shooting percentages for the rest of the 2008 data set.

A picture containing timeline

Description automatically generated

Figure : Heatmap of 2008 Players without Yao Ming

Therefore, the original 2008 data was remapped with Yao Ming removed to better visualize the relative 3PP performance of the top players of the time. In the 08 dataset, we can see that amongst top players in the league, 3PP is at a similarly high level across the board, with a few standout players. This would indicate that being a good shooter, especially at the 3-point line, was an important part of a primary scorer's game. This is further corroborated by many players also having a high free throw % (FTP). There are a few data points to specifically note, particularly the very dark blue and white spots on the heatmap. These extremely high and low 3PP stats reflect the playstyle of the most successful power forwards and centers of the league, that took few to no 3-point shots, causing that stat line to show either extremely high or low percentages. These players include Pau Gasol and Yao Ming with extremely high percentages, and Shaquille O'neal, Tim Duncan, and Dwight Howard with extremely low percentages. These players typically found their scoring success with high-percentage shots inside of the 3-point line using their size and strength, which is supported by many of these players also having extremely high Field Goal Percentages (FGP). Furthermore, these players were successful due to other aspects of the game than scoring, indicated by many players with extremely low 3PP also having unusually high Rebounding and Block statistics.

We can also discuss certain players that display deeply shaded blocks on the heatmap, indicating their unusually high performance in certain areas of the game. One such player is Chris Paul, who is close to the top of the Name axis and has very dark blocks in the Assist and Steal categories. This indicates his role as a crafty and strong playmaker, finding success not just as a scorer by himself, but making plays happen on both sides of the court. Another notable player is Dwight Howard, who notably has virtually no value in the deep shooting categories. However, it is clear he found success in other ways, shown by his incredibly high value in all rebounding categories, as well as in blocking and free throw attempts. This indicates his success as a defensive player and with plays at the rim earning him many free throws from other players fouling him in the act of scoring.

Other players that stand out include Kevin Martin, who is appears to be a good shooter in general but was the best in the league at shooting free throws. Another is Deron Williams, with assists at about the same level as Chris Paul, and Stephen Jackson, with the most turnovers in the league. Lastly, Corey Maggette has the most Personal Fouls called against him by far, which upon further investigation, reflects his ability/style to draw fouls while scoring and create points from his solid free throw shooting.

**2018**

**A picture containing background pattern

Description automatically generated**

Figure : Heatmap of Top 50 2018 Players

As for the 2018 data, we see fewer clear patterns and players that easily stand out from the rest across the board. This may be an indicator that the way the game is played has shifted, with many players developing other aspects of their skill set especially as 3-point shooting has become more emphasized.

While there are fewer clear stand-out players, we still do see some players excelling in certain areas. Firstly, we can see that James Harden leads the league in Points by a noticeably larger margin than was seen in the 2008 data, with the shading in the PTS column dropping off much more sharply. This, coupled with his league leading free throw percentage (FTP) indicates his position as a top scorer in the league with the ability to gain many free throw attempts at the line from fouls drawn on his plays.

Another notable player is Stephen Curry, with excellent free throw, field goal, and 3-point shooting, displaying his season-leading abilities as one of the best-shooting point guards the league has ever seen.

On the other side, we can also see top inside scorers in Giannis Antetokounmpo, with leading stats in the 2P-shooting %, Field Goal %, and rebounding, displaying his dominance as the leading "big man" in the league.

Other notable individuals are as follows:

-Highest Percentage 3-point shooters: Buddy Hield, Bojan Bogdanovic, and Danilo Gallinari (who also appears to have the highest offensive rating in the league)

- Highest Percentage Field Goal scorers: Giannis Antetokounmpo, Stephen Curry, JaKarr Sampson, and John Collins

- Best Free throw Shooters: Stephen Curry, Damian Lilliard, Danilo Galinari, and JJ Redick

- Best Assister: Russell Westbrook

- Best Stealers: Jimmy Butler (by far), and Paul George

- Best Blockers: Anthony Davis (by far), Giannis Antetokounmpo, Joel Embiid, and Karl-Anthony Towns

- Best Rebounders: Joel Embiid, Giannis Antetokounmpo, and Karl-Anthony Towns

- Most Turnovers: Trae Young

We do see certain players appear in both the 2008 and 2018 datasets, namely Kevin Durant, Lebron James, and LaMarcus Aldrige. Within the top 50 players from each dataset, these are the only apparent players that are still playing in the NBA 10 years later. After further investigation, all three of these players were drafted within a few years of the 2008 season and thus were very young when the data was collected.

When comparing players that appear in both data sets, age can be a factor in predicting player success in that players who excelled at a high level early on in their career also tend to find success much later in their careers as well, barring serious injury. However, it is difficult to predict player success off the provided data since virtually all top players in the league in 2008 are not present in 2018, as most were at the height of their careers in 2008 and retired by the time the 2018 season arrived. Exceptions to this are the players stated above.

**Conclusions:**

The most successful players in the league can most simply be identified by ranking players by how many points they scored. However, this is not a complete picture, and players found their success in different specific aspects of the game. From looking at both 2008 and 2018 data, we can see that these players were able to find opportunities to have impact in their teams typically from either being excellent shooters, defenders (blocks and rebounds), playmakers, and/or insider scorers. Furthermore, we can see that typically, excelling in any category does seem to have a link to performing better in certain areas and worse in others, as plays tend to follow a particular playstyle based off of their speed, size, and ability to score, often at the cost of other areas of the game.

For example, top shooters such as Stephen Curry, Buddy Hield, and Tobias Harris tend to lack in areas such as blocking or rebounding but find great success shooting the ball, whether at the free throw line or from the 3-point line. On the other end, we can also see the "big men" of the league finding success with high defensive statistics such as blocks and rebounds, sometimes paired with high field goal percentages (especially in 2008), indicative of them usually preferring to score with high percentage shots inside of the 3-point line.

We can also see which players found great success as playmakers, displayed by their high assist statistics that is usually linked with high turnovers. While at first glance having high turnovers may seem to be a very negative statistic, this is simply indicative of how much of their team's offense flows through these players, as they primarily decide where the ball moves as the play develops. Since these players handle the ball so often, it is sensible that even very successful assisters will have a high turnover rate by virtue of handling the ball so often.

Dendrograms of the above heatmaps may be viewed below as well (Names are represented by colors but can be seen from within RStudio):

Chart

Description automatically generated

Figure : Dendrogram of 2008 Player Data without Yao Ming

Chart

Description automatically generated

Figure : Dendrogram of 2018 Player Data

data set 2.

(Devin)

Question 1.

**Use PCA to reduce dimensions. How many components do you need to keep to reproduce the digits reasonably well? what is your final matrix?**

To capture ~94% of the variance, 75 PCs is enough, and based on plots of the cumulative variance, the elbow at which diminishing returns for including additional PCs starts about here. The drawn images are recognizable at this level, and even at lower thresholds around 20~30 PCs the images can be visually interpreted. However, the background is not correctly reflected as completely white, and the label assignments are not correct even at 75 PCs or higher, despite cumulative variance totaling ~99%. To reach the highest level of accuracy at which the background and labels are correct, 567 PCs are needed. This amount is still 75% of the original number of components, which isn’t an excellent reduction of dimensions. Excluding the labels could potentially improve the dimensionality reduction, though they are a single column with fairly low variance.

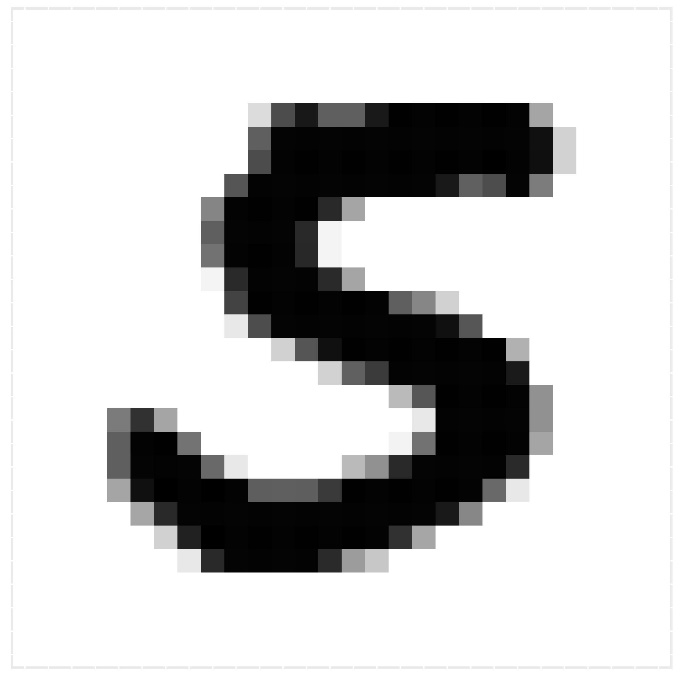


Figure 6 Digit drawn with original data for reference

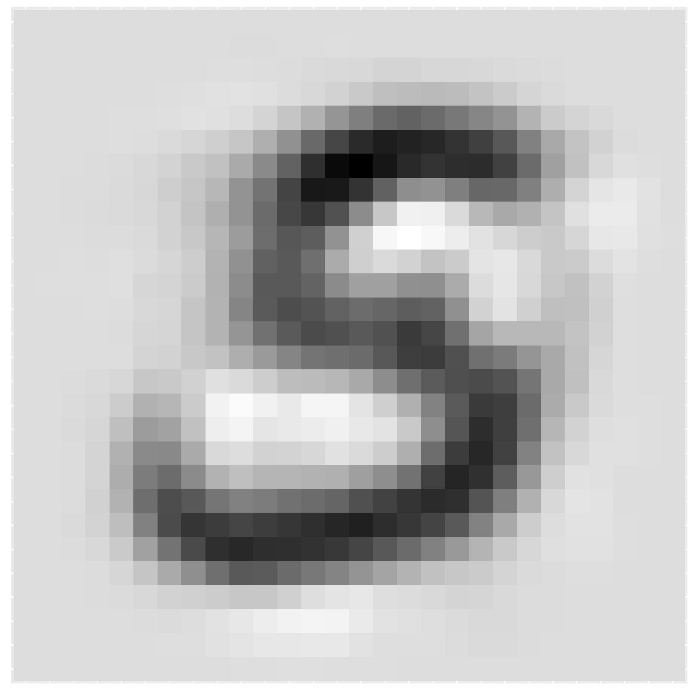
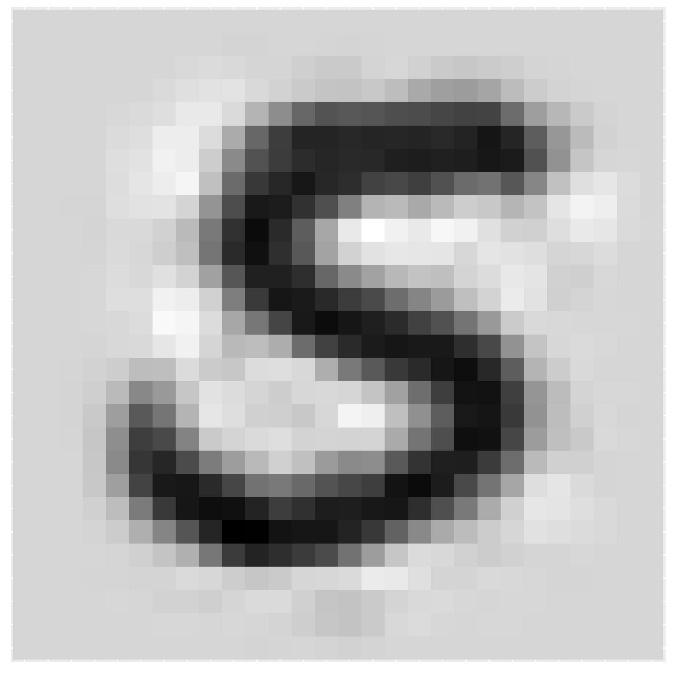


Figure 7Digit drawn with 20 PCs

Figure 8Digit drawn with 75 PCs

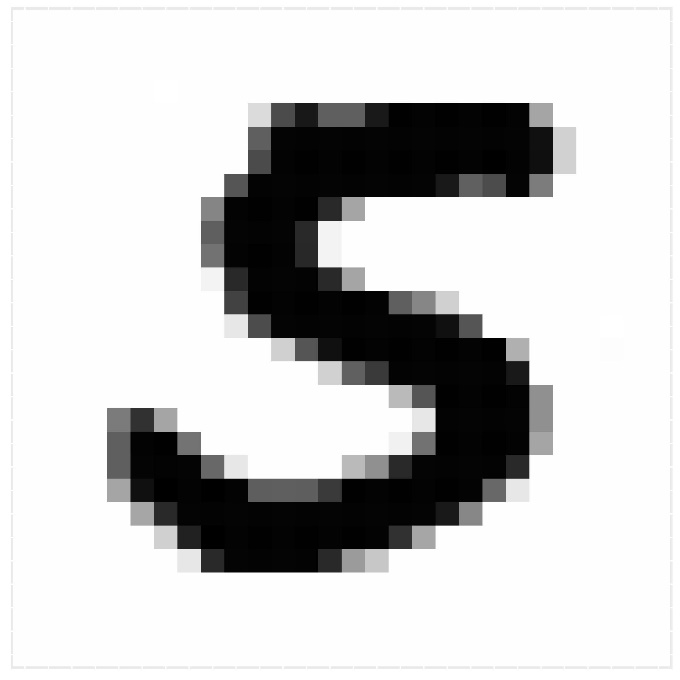


Figure 9Digit drawn with 567 PCs

## Question 2.

**Draw a tree of the pixels, and see if you can explain the results based on geometry of the pixels (how far apart are they in the 2-d space). Try to Explain the PCA results in light of this.**

By creating the dendrogram, we can see that the pixels with little to no data in them (mostly the edges of the images) cluster together and make up about 25% of the data. This is consistent with the above finding that 75% of the PCs were required to recreate the images.



## Question 3.

**Can you use some of the tools you have learnt to build a classifier,so if you get a new set of pixels you can predict what is in the picture. This is a the start of a real project, but you don't have all the tools (such as neural networks) which might be more suited for this task. Split your dataset into two (a training set and a test set), build your classifier and figure out how well it does with the test data in predicting the digits. Define the sensitivity and specificity of your classifier. How well does it recognize your own handwriting? (make sure your handwriting is not in the training set)**

This will be addressed using three different levels of algorithmic complexity.

3.a. Machine learning algorithms/neural networks are the obvious choice for this task, but first we will try to classify by simply using the average for each digit and correlating against that. Using this method is about 80% accurate. When comparing to the handwriting samples I created, it was limited by the fact that there was only a single comparison, between the average and the sample for each digit. It was able to identify five of the digits correctly.

Sample classification based on correlation:

A picture containing logo

Description automatically generated

Results of first 10 samples:

Table

Description automatically generated



3.b. The results of the correlative approach seemed insufficient, so using a machine learning model with PCA was the next step. By reducing dimensions and using LDA, we hope to classify each digit using a smaller number of the PC values instead of all 784 pixel columns. However, the performance of this model is poor despite various attempts to optimize, including using up to 500 PCs. The specificity and recall (the term sensitivity only applies to binary classification) were both far below acceptable standards, approximately between 0 and 0.2 depending on the digit being evaluated.



3.c. The PCA model was disappointing as well. Thus, a more proper machine learning algorithm seems to be required. We next evaluate a random forest model. This model uses much less time and lines of code, and has an accuracy of 95% when applied to the test set. However, it is completely unable to identify more than two of the handwritten samples. When applied to the averages for each digit, it can correctly identify 4 out of the 10. The failure to identify the handwriting samples may be due to the scale of the images in both position and intensity - the amount of whitespace outside the writing is not exactly the same, and the intensity of the black color is less in the samples. It may perform better if each image in both sets had any row or column in the 28x28 image with a maximum value below a certain threshold removed, and with the color normalized so that the darkest black part is always equal to the maximum.

Test data accuracy (First 20 predictions):

Table

Description automatically generated

Handwritten digit accuracy:

Table

Description automatically generated with medium confidence

As a final observation, the basic correlative approach performed surprisingly well compared to more advanced algorithmic techniques.

## Question 4.

**You can try simple things like take average of all data for each number and then take a "dot" product with your test set, and identify the pixels. This might work, maybe for some digits, and not others.**

Here we will see which of the digit averages is best classified by the Random Forest model. This model is only about to correctly identify 4 of the 10 digits.

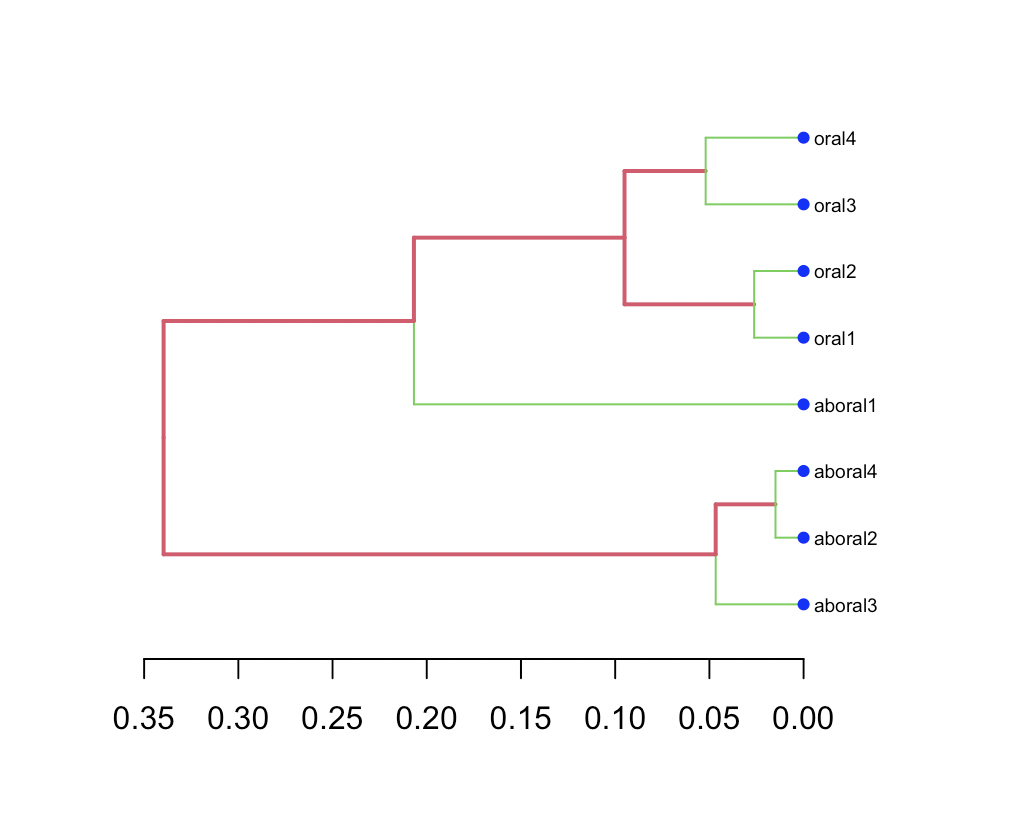
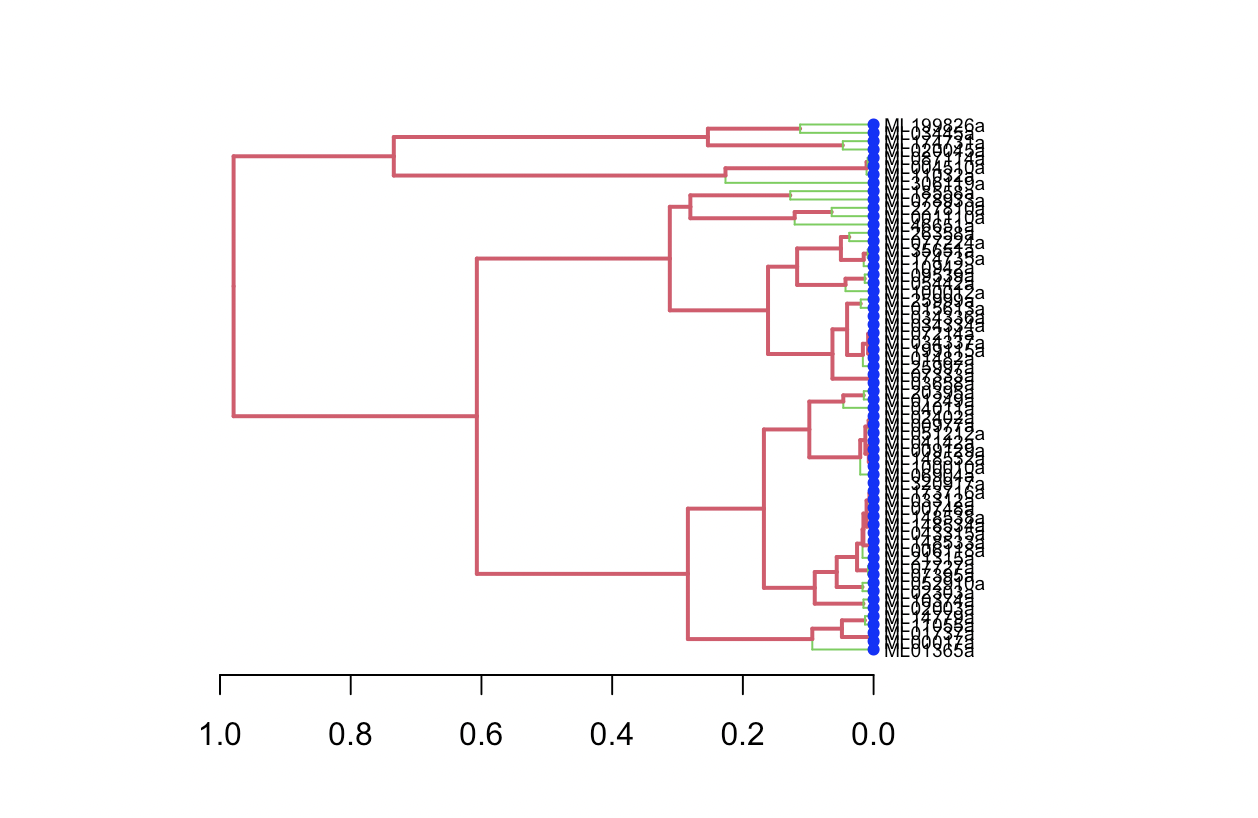
Average digit accuracy:

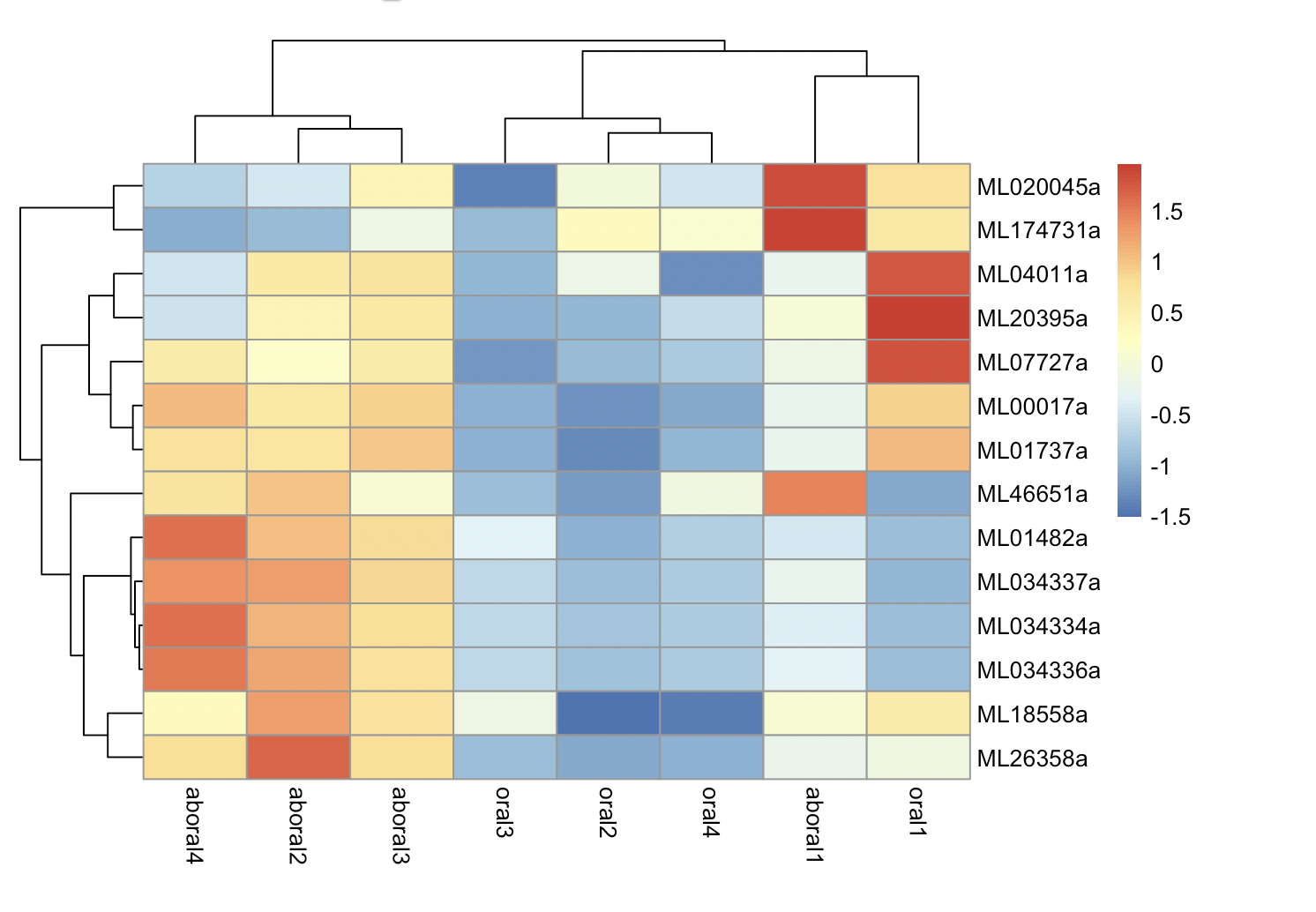
Table

Description automatically generated

# Data set 3:

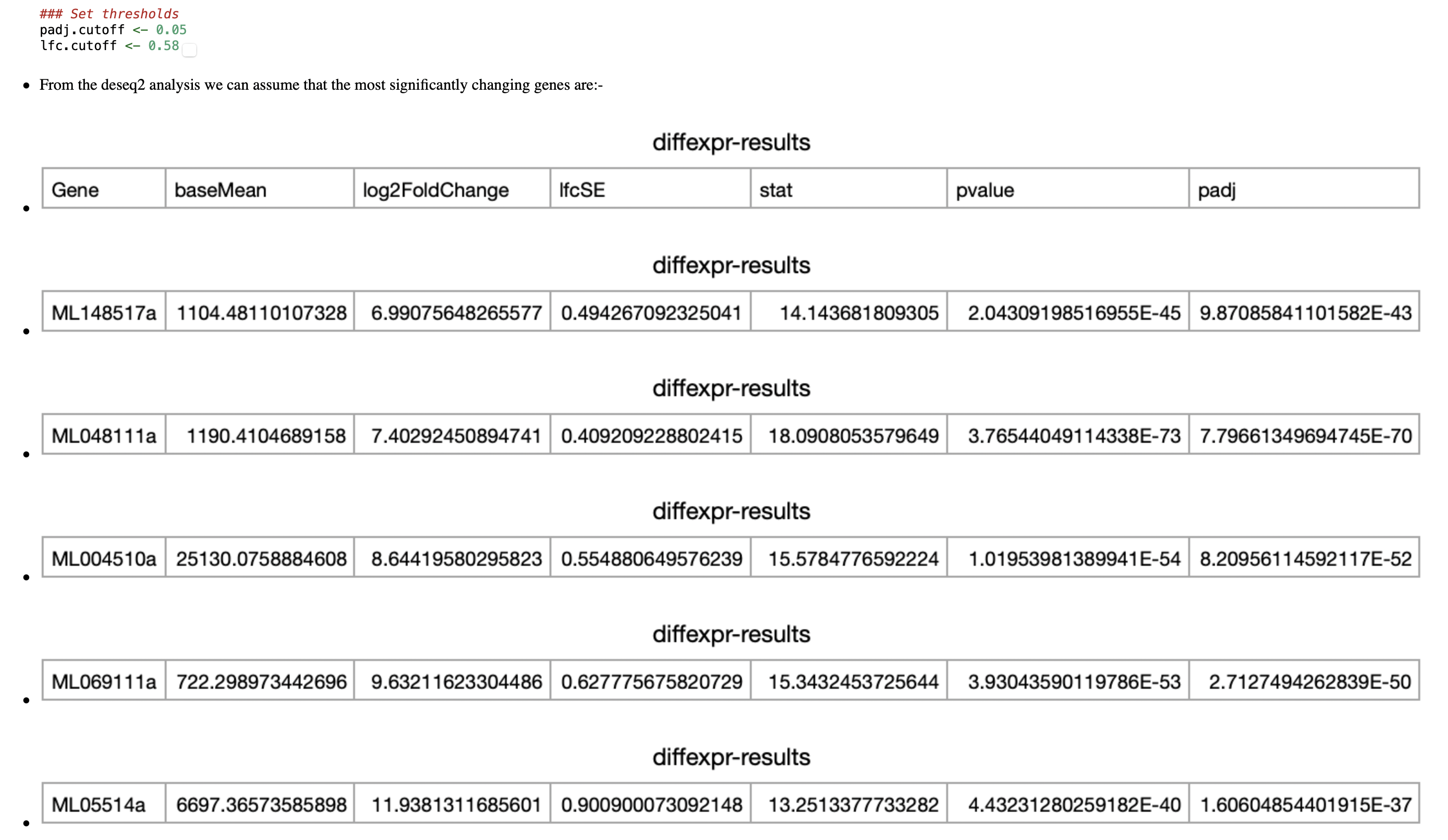
(Mudra)

**1) Build hierarchical trees based on the columns and for the rows (exclude rows that are "low" expression)**

**2) Draw a heat map of the expression data**

**3) Use DESeq2 to analyse this data**

**a. Which are the most significantly changing genes in this dataset?**

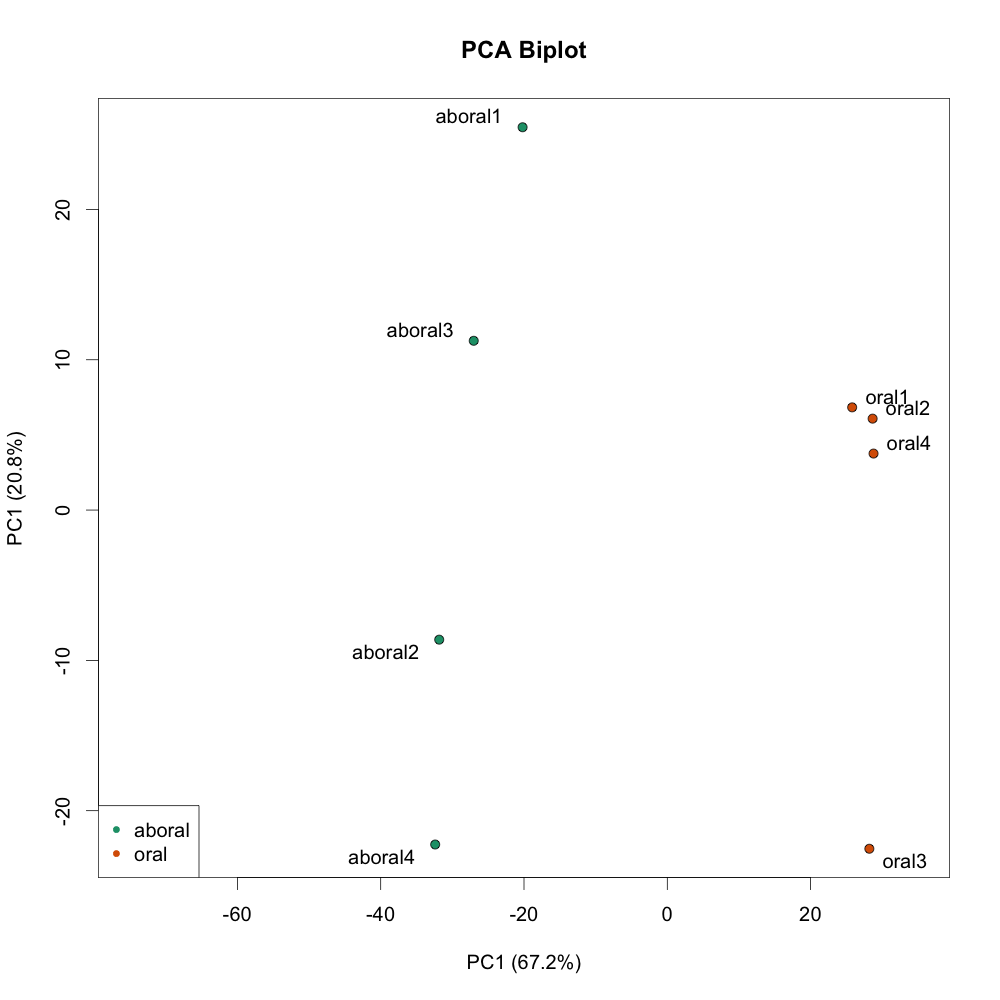
There are couple of columns that we can look at from the "diffexpr-results.csv" file to indicate which genes are significantly changing. Looking at the log2FoldChange column a positive fold change value indicates an increase of expression, while a negative fold change indicates a decrease in expression. The table is organized by the adjusted p-value column which indicates whether the gene analyzed is likely to be differentially expressed in that comparison. This applies to each gene individually, assuming that the gene was tested on its own without consideration that all other genes were also tested. 2129 genes are significant with an adjusted p-value <0.05. What we noticed is that the FDR threshold on its own doesn’t appear to be reducing the number of significant genes. With large significant gene lists it can be hard to extract meaningful biological relevance. To help increase stringency, one can also add a fold change threshold.

**How consistent are these results with the analysis you did in the midterm project?**

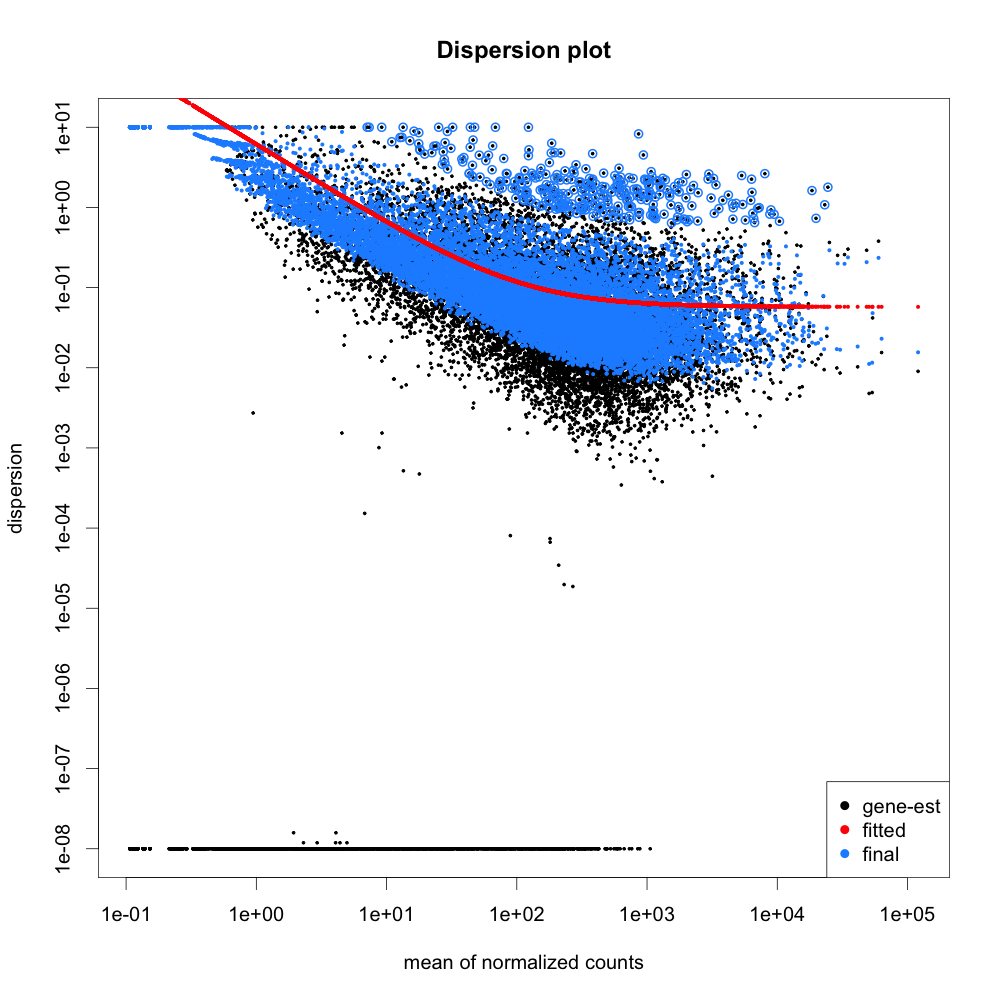
The data varies a bit from the midterm project however some data is similar. For example, ML034332a was classified as a down regulated gene in the midterm project and based on the Deseq2 analysis (MA Plot) it is an outlier gene. Also, when the genes were organized via expression for column vs row basis in the midterm project it showed 'ML004510a' to be in the top 5 varied gene however this gene is also another example of an outlier.

**What else can you say about the data in terms of consistency, and the results that you find from your analyses. The question is open-ended, think of this as your experiment, you need to write a paper based on this data so you have to figure out what kind of "story" you can tell based on this.**

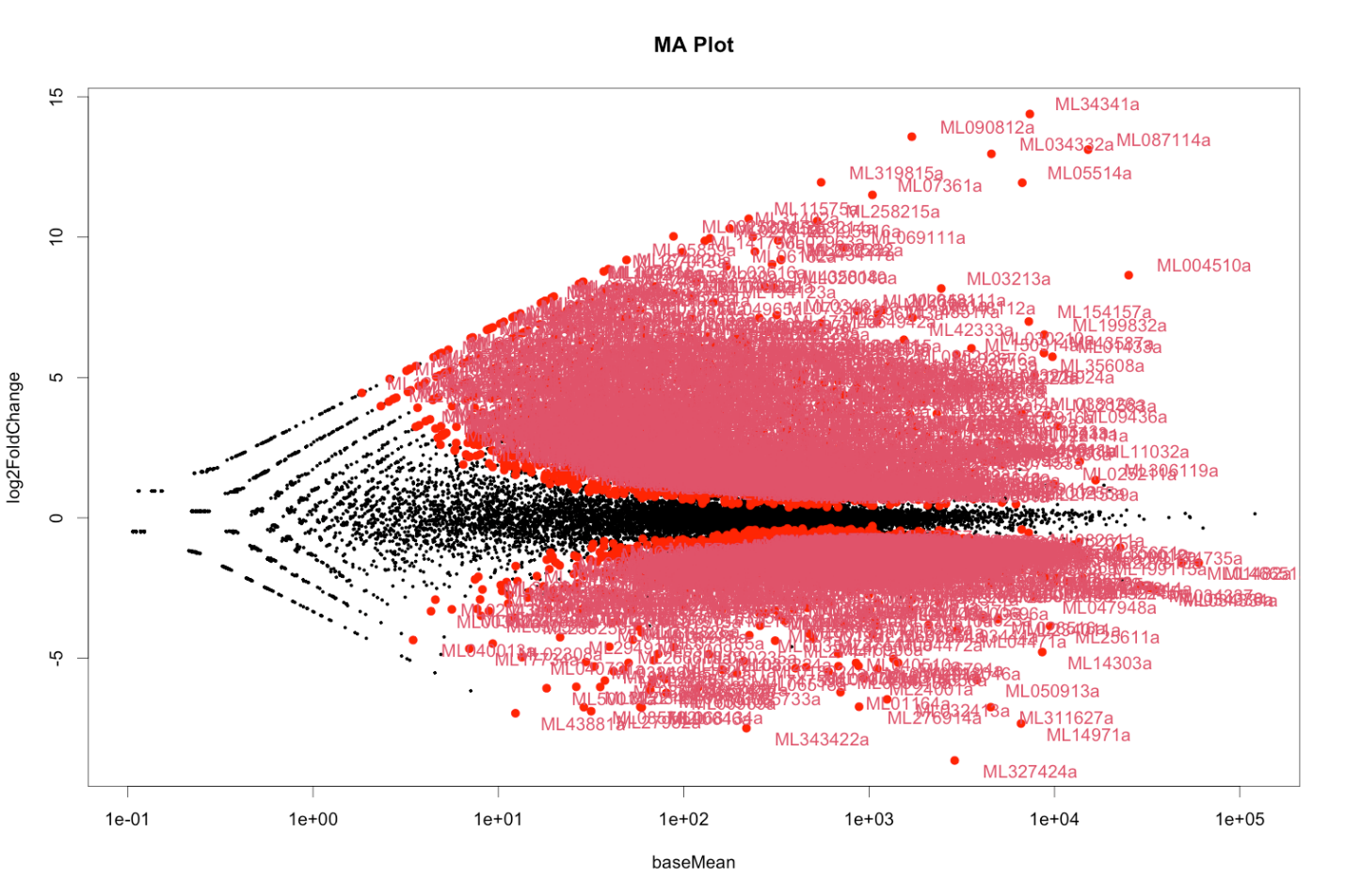
Principal Components Analysis (PCA) is a dimension reduction and visualization technique that is here used to project the multivariate data vector of each sample into a two-dimensional plot, such that the spatial arrangement of the points in the plot reflects the overall data (dis)similarity between the samples. The majority of variation between the samples can be summarized by the first principal component, which is shown on the x-axis. The second principal component summarizes the residual variation that isn\'t explained by PC1. PC2 is shown on the y-axis. The percentage of the global variation explained by each principal component is given in the axis labels. In a two-condition scenario (e.g., ABORAL vs ORAL), you might expect PC1 to separate the two experimental conditions, so for example, having all the controls on the left and all experimental samples on the right (or vice versa - the units and directionality isn't important). The secondary axis may separate other aspects of the design - cell line, time point, etc. Very often the experimental design is reflected in the PCA plot, and in this case, it is.



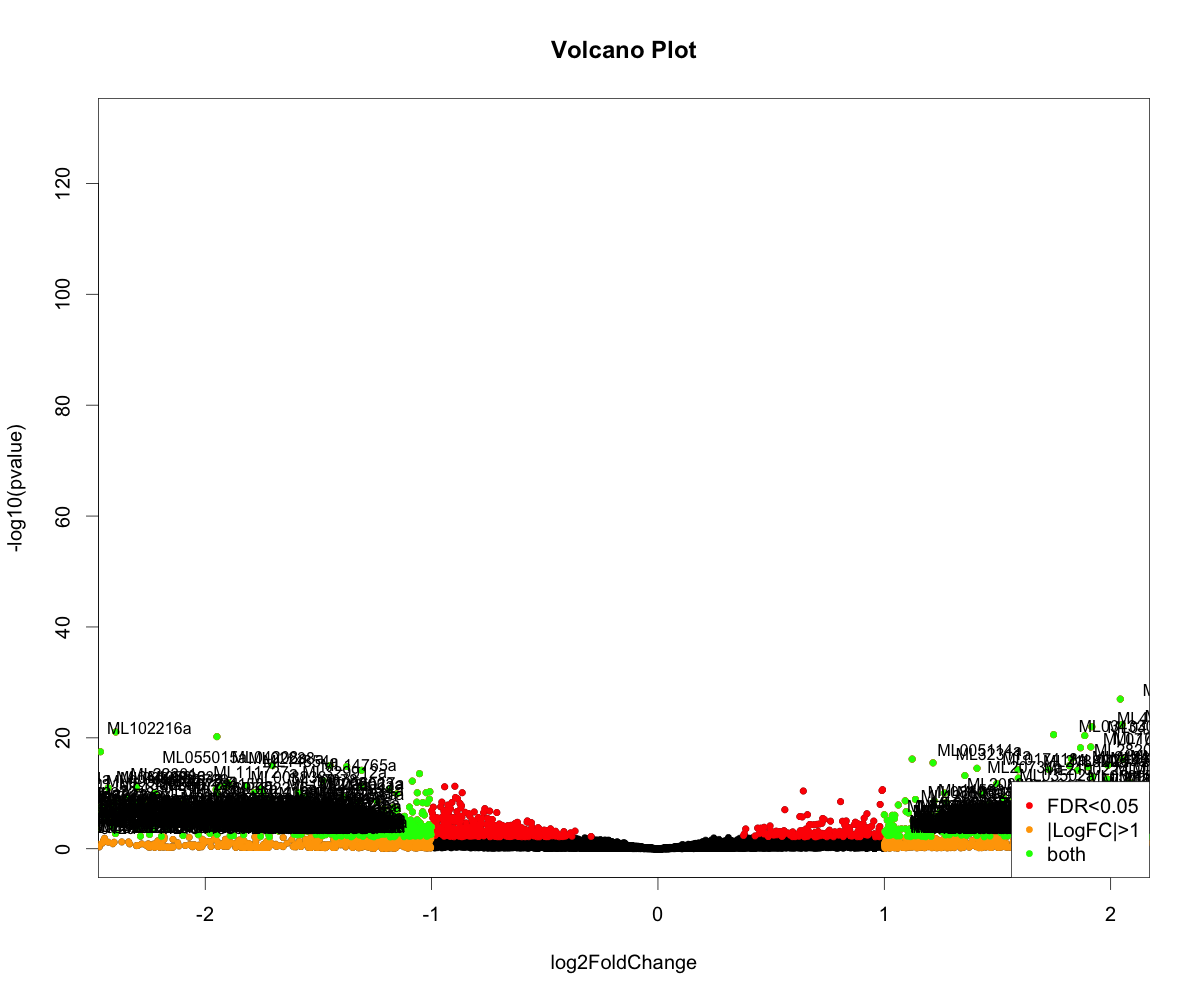
The reason behind fitting a curve to the data is that different genes will have different scales of biological variability, but, over all genes, there will be a distribution of reasonable estimates of dispersion. This curve is displayed as a red line in the figure below, which plots the estimate for the expected dispersion value for genes of a given expression strength. Each black dot is a gene with an associated mean expression level and maximum likelihood estimation (MLE) of the dispersion. This plot is a great representation of examining data after deseq2 analysis. You expect your data to generally scatter around the curve, with the dispersion decreasing with increasing mean expression levels. If you see a cloud or different shapes, then you might want to explore your data more to see if you have contamination or outlier samples.



An MA plot shows the average expression on the X-axis and the log fold change on the y-axis. This MA plot shows a high number of data points falling above the one threshold on the y-axis indicating a more significant number of genes being upregulated, while more below −1 would indicate high levels of downregulation in genes.



A volcano plot shows the log fold change on the X-axis, and the −log10−log10 of the p-value on the Y-axis (the more significant the p-value, the larger the −log10−log10 of that value will be). Looking at the x axis we can see that all the genes on the right of the 0 (positive side) are all positively expressed whereas the ones plotted on the left of 0 are negatively expressed.



**What is the most interesting pathway or gene that is responding in this study?**

ML01051a seems to show all the properties for being an outlier, however it was in the top 5 of high expressed genes.