AN APPLICATION OF THE ONE-CHANNEL AFFYMETRIX MICROARRAY ANALYSIS

ELİF DOĞAN DAR

STAT 730, 02.06.2017

1) INTRODUCTION

The smallest structural and functional unit of any organism is a cell. It consists of many subunits, organelles, including nucleus which controls all functions of the cell by regulating gene expression. Nucleus has highly packed DNA which contains all the information needed to keep the cell running. DNA has a double-stranded helix structure where there are complementary chemical bases, adenine (A), guanine (G), cytosine (C), and thymine (T), on both sides. Adenine is paired with thymine and guanine is paired with cytosine. This pairing principle is being used to extract information about the cell through experiments like microarray experiments. These lined up bases create genes which act as instructions to make molecules called proteins, which are responsible for the functions in a cell. In humans, genes vary in size from a few hundred DNA bases to more than 2 million base and humans have between 20,000 and 25,000 genes. All cells have the same DNA in their nucleus but which genes are active changes from cell to cell. While tracking the genetic diseases this fact is being used. Malfunction of some genes, caused by a mutation or alteration, cause cancer cells to grow faster than they should. If we can detect which genes are responsible from that, it would help us to detect whether a person has this cancer or not. Also, it would help us to understand the mechanisms behind and maybe this way new approaches for the cure can be suggested. To be able to detect which genes are active in a sick and a healthy cell, one can use microarray analysis. To be able to understand the idea behind microarray we need to understand the transportation of the genetic information.

Information in the DNA has to be transported inside the cell. To achieve this, the cell produces messenger RNA(mRNA), DNA helix opens and some part of it which is going to be used, genes, are being copied. This way one strand of this gene information is being copied as mRNA and mRNA is being sent to the outside of the nucleus to be used. Therefore, at any given time cytoplasm has a big amount of mRNA. If we can see this information carried by mRNAs, then we would know which genes are active in this cell. We do not have the technology to simply look at a cell and detect how many mRNAs are in it and which genetic information they carry. Therefore, microarray technology has been developed. Thanks to this microarray set up we can analyze and decide which genes are active in a cell. Microarray chips consist of microscopic DNA spots, namely probes, each representing a gene, attached to a solid surface. When we put mRNAs taken from a cell on this microchip, they will bind to their complementary parts on the chip. So, if we can detect the amount of mRNAs on a particular probe we will know how much active this gene is in that cell. Since simple counting is impossible for now we attach a fluorescent dye to mRNAs before. By measuring the amount of color emitted by the array we will be able to say the amount of mRNA in the cell. A more intense signal will mean a more active gene.

2) PREPARATION OF THE DATA

Preparation of mRNAs which we will be using in the experiment consists of many stages. First, we will collect tissue samples from healthy and sick individuals. We dissolve this tissue sample in a mixture of various organic solvents, this ways DNA, proteins, other cell components and RNAs will be dissolved. Then we will put this mixture in the tube into a centrifuge machine. As the centrifuge spins with a high speed, heavier ingredients will descend and lighter RNA will be concentrated on top of the tube. When we separate this upper part from the rest we get RNAs, however, this process doesn't give a perfect separation. Therefore, we repeat this purification many times until an acceptable ratio of RNA is being reached.

Unfortunately, this new sample also includes ribosomal and transfer RNA together with mRNA and only mRNA reflects gene expression. Therefore, we need to separate mRNA from the others. Among them, only mRNA always end in a sequence of adenines, known as a “poly-A tail.” We will use beads with poly-T tails attached to them. When we make our sample pass through a tube full of these beads, only mRNAs with poly-A tails will get attached to them. Now we have to separate the beads from the mRNAs.

The only thing which keeps beads and mRNA together is the chemical bonds among T and A tails. These bonds get highly affected by chemical conditions such as salt concentration and pH. When we wash these beads with a buffer which has proper conditions, chemical bonds will break and we will be left out with mRNAs only. Now as mentioned before we should dye mRNAs, this can be done separately to both cell types by using two microarrays or we can use only one microarray for both which will gain us time and resources. To achieve this, we change mRNA molecules with dyed cDNA molecules. Dyed bases are added to the mixture of mRNA together with poly-T primers and reverse transcriptase. This way dyed cDNAs will make a complementary strand for mRNA and mRNAs are degraded. Now we have more stable cDNA molecules instead of easily degradable mRNA. Let's say we dyed mRNAs taken from cancer tissues with red and healthy tissues with green. When we examine intensities, red spots will indicate genes which are active in cancer cells and inactive in healthy cells, green spots will indicate the opposite and yellow ones will be active genes in both cell types.

After preparation of cDNAs, we will also need a microarray chip. To prepare a chip, more advanced and complicated technologies needed. Gene sequences should be known in advance, they should be produced artificially and put on different probes. Since it is a difficult task to produce for most of the labs, there are companies which produce and sell these microarray chips. One of the well-known brands among them is Affymetrix and we will be using data obtained from Affymetrix in this project. Dyed cDNAs will be added to this chip and light intensities will be recorded. This will be the data we obtained from the experiment. It is a matrix with genes in the rows, different cell types as columns and light intensities as entries.

2) NORMALIZATION

There are two sources of variation in microarray data, random and systematic variations. Random errors are related to the nature of the data and cannot be decreased, where systematic errors have many reasons related to the experiment and can be decreased by certain methods, this data cleaning process is called normalization but before normalization first, there are some control mechanisms that can be checked.

Control Mechanisms: While preparing the probes on the microarray, some unrelated genes, whose intensities known in advance, are being attached to some of them. If the measured intensity is different than what we expect, there might be a systematic error in the experiment. However, this difference might be caused by random error, to make sure it is systematic we increase the number of these spots, distribute all over the microarray and check all of them. Also, we prepare some empty probes on the array, intensities of these probes will give us the least possible intensities on the chip. Finally, there are bright spots, “landing lights”, in the corner of the array that help in the alignment of the array while scanning. Another control mechanism is the investigation of the outliers, because of some damages on the array, wrong applications throughout the experiment or in the scanning process we might get faulty signals. Intensity of one probe might be brighter or darker than it should be. On the array, there are many probes representing the same gene. We compare intensities of probes representing the same gene and flag the ones with clearly different values as missing. Then either we exclude this values from data or replace them with some different imputation techniques.

Adjustment of scale: In this microarray experiment, what we are trying to do is counting the number of mRNAs of the genes. However, what we actually get is optical intensities of the dye molecules attached to them. The relation between these two is only partially linear, especially in high and low-intensity values there is too much distortion of linearity. There are some proposed methods to get rid of nonlinearity on the tails. It has been shown that some of the high-intensity values are reaching the most possible values. They showed that by using some method with maximum likelihood techniques and several parametric assumptions these signals can be recovered. For low-intensity values, other problems exist, the relative error increases for these spots. Log transformation can be applied to solve these problems. In this project, we will be using logarithmic scale, there will be no loss of information by log scaling and fold changes will result in same absolute change which is not possible in the original scale. Except for background normalization, log transformation works very well in all normalization techniques.

Sources of variation: Repeated experiments never give the same results. There are many sources of variations in the experiment on different levels. It can be caused by cDNA, microarray production, hybridization process and scanning. There can be differences between samples taking from the same subject, differences in the conditions of the time taken, variation in mRNA extraction methods or transcription. Labeling efficiencies change over time or for different dyes, also some dyes are more sensitive than the others. ON the production of the microarray if the is the difference in sequence length of the DNAs on probes, it affects the intensity. There are variations on different chip batches and chemical probe attachment levels to the chip. There might be inequalities while applying cDNA to the chip or variations in washing efficiencies of non-hybridized cDNA off the slide. There might be differences in temperature, experimenter or time of the day. And finally, there can be scanner variations caused by the usage of different scanners, different spot-finding software or different grid alignments. By applying different normalization techniques, some of these variations can be decreased. There are four major stages of normalization as follows;

Spatial Normalization: When we look at the optical image of the samples, there might be some unequal color distribution. Sometimes it is expected if related genes are put in the spots near each other or control spots are in the same area. However, that is not the case most of the time and it indicates some systematic error; unequal distribution of the cDNAs, an unequal wash of the chip or scanning errors. There are many methods to get rid of this systematic bias.

Background Normalization: Background signals are faulty signals caused by some wrong hybridizations. Some of the mRNAs can bind to non-probe areas, some can bind to some probes although there is no match and sometimes they can bind to some areas where there is only a partial match. The proportion of faulty signals to the true intensities cannot be detected since as data we have only total intensities. Affymetrix developed a method to overcome this problem. For every probe (called as perfect match(PM) probes), they added another probe whose genes have a different 13th base (called as mismatch(MM) probes). The intensity of the MM probes gives the amount of faulty signals. However, there are some problems with this approach, even though PM and MM probes are very close to each other the amount of mishybridization might not be equal to them. Changing only one base will increase the probability of having mishybridization because lots of matching cDNAs will partially match to the MM probe. Since this is the only option we have for background normalization we will use this approach. There are several methods to apply for background normalization, we will use RMA(Robust microarray analysis) method. This method is non-parametric, so there will be no need for distribution assumption. Also, it is a robust technique, it is not highly affected by outliers.

Dye Effect Normalization: In two-channel microarray experiments, mRNA molecules of different cell types are labeled with two different dye molecules, Cy3 and Cy5. However, sensitivities, sizes, and reaction to photobleaching of these two dyes are different than each other. There, this dye effect should be corrected. One of the proposed methods suggests swapping the dye molecules, repeat the experiment again and take averages of two intensity values. Since our data is one color, there is no need for dye effect normalization process.

Quantile Normalization: In microarray experiments, the distributions of probe intensities under the same condition are expected to be the same. But this is not the case in most the microarray data. While preparing hybridization samples or in the scanning process, there are many stages that can cause a big amount of variation. Therefore, first, we should put all arrays on the same scale by doing quantile normalization and then do comparisons. There are two different stages for quantile normalization, within replication normalization and across condition normalization.

Within-replication normalization: It has been said that distributions of probe intensities under the same condition are supposed to be same. To achieve this, a proposed method transforms all replicates onto the same scale. While doing that method takes only ranks of the observations, therefore it works well even with the non-normal distributions. First, we order all entries in the arrays of some specific condition from smallest to largest. Let's say we have n different arrays belong to the same condition. We look at first n entries and replace them with the mean of these n entries in the original data, we do the same for the second n entries, third n entries and so on. This way we put all arrays of the same condition on the same scale.

Across-condition normalization: The same problem arises across conditions. To make them comparable we should put them on the same scale. Although it is reasonable to expect that probe densities will follow the same density for the same condition, it might not be the case for arrays under different conditions. Different conditions might affect gene expressions. Therefore, one should be more careful while doing across condition normalization, over normalization shouldn't take place. To achieve this, we will choose invariant genes, which are known beforehand, under these different conditions, apply quantile normalization to this set of genes and interpolate the remaining ones. However, interpolation only works if remaining entries are in the range of the invariant set. This is achieved by forcing genes with minimum and maximum intensity values to be in the invariant set.

4)APPLICATION ON DATA

We have Affymetrix one-channel microarray data. Eight samples have been collected; two adult brain samples, two adult liver samples, two fetal brain samples and two fetal liver samples. We would like to know which genes are active in one group while inactive in another group. First, we read our data in CEL files into R by using Bioconductor and affy packages by ReadAffy() command. First, the image of the microarray data is checked for being aware of any spatial effect. We will use image() function to get this image. As you can see in *Figure 1.1*, left corner of the last array seems like a problematic region. Other than that region, there does not seem to be any problem.

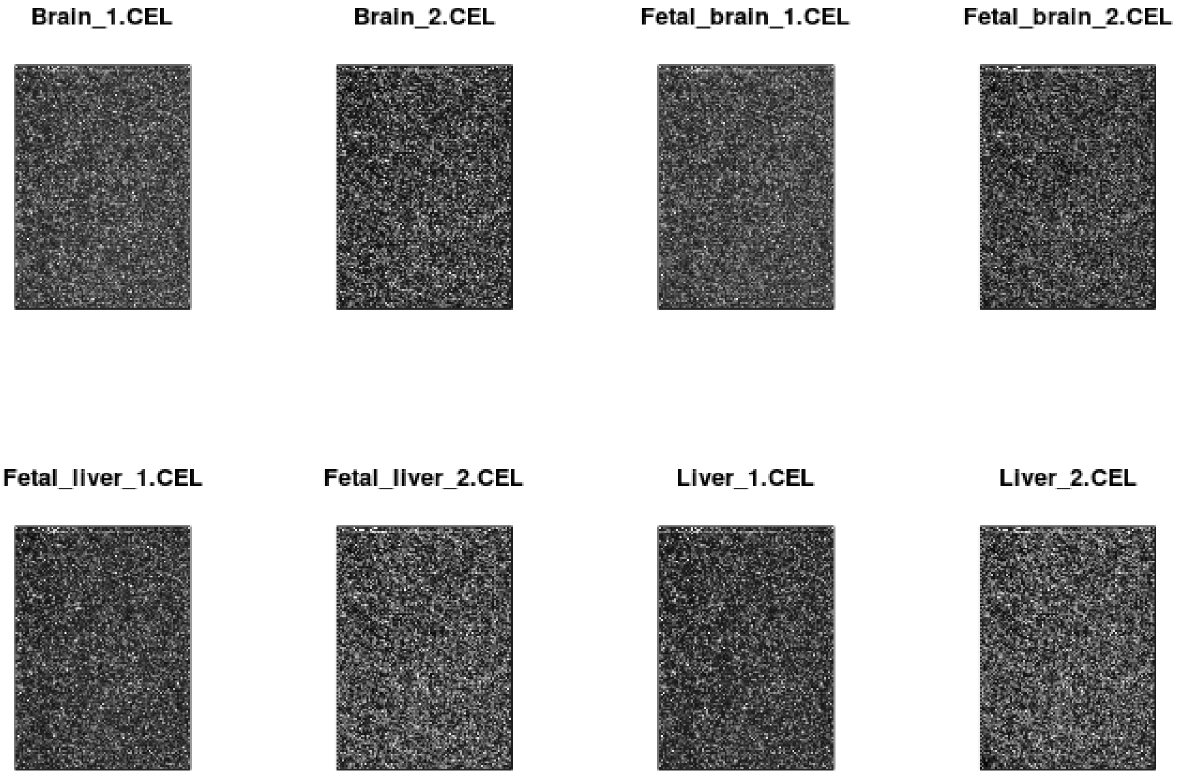

*Figure1**Error! No text of specified style in document.**.1: Images of raw microarray data*

Also, we can check densities of the arrays. We will plot these densities by using plotDensity.AffyBatch () function. As you can see in *Figure 1.2*, densities of these 8 arrays are different from each other. There will be need for quantile normalization.

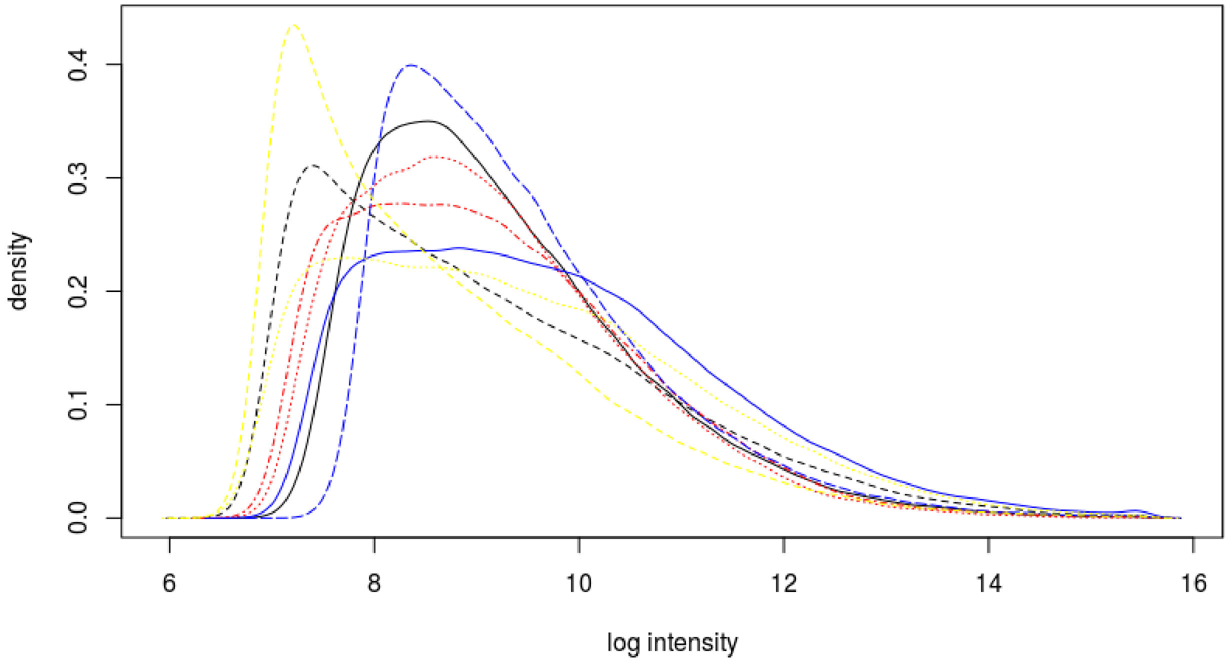

Figure1.2: Density plot of unnormalized data

Now let’s look at the boxplots to get a clearer understanding of data, *Figure 1.3*. There is not much difference in means and variation between fetal brain samples. However, there is an obvious difference in mean of liver samples and difference in variation between fetal liver samples and brain samples. It is clearer now that to be able to compare these arrays some quantile normalization has to be applied.

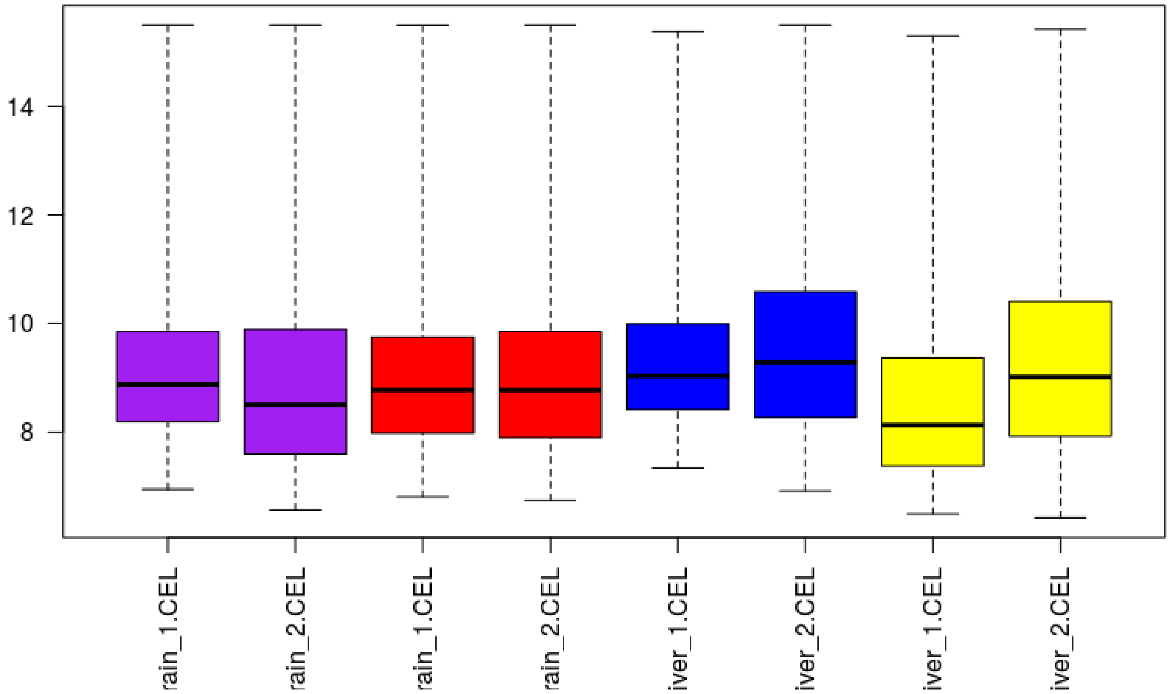

Figure1.3: Boxplot of raw data

Later we check some of the MA plots for this raw data by using mva.pairs() function. As you can see there is a pattern in the plot although the differences had to be randomly distributed. We will check later whether these problems could be fixed or not.

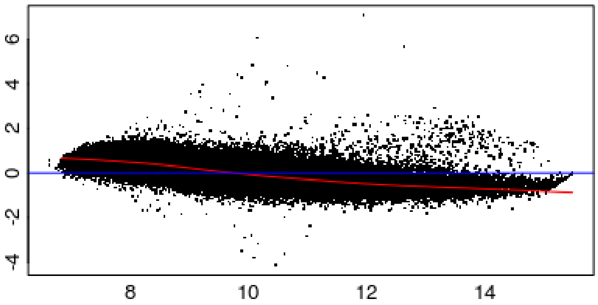

Figure1.4: MVA plot for brain sample 1 and brain sample 2

Now we will check some quality measures by using qc() function. When we plot this function, we get a figure shown in *Figure 1.5*. The first quality measure is the average intensities of the background probes on each array, these values for different arrays should be comparable. The second quality measure is the scale factors: factors used to equalize the mean intensities of the arrays, they should be within 3-fold of each other. The third quality measure is the percent present calls: the percentage of spots that generate a significant signal (significantly higher than background) according to the Affymetrix detection algorithm. The present calls should be similar especially among replicates. Very low values (<20%) are a possible indicator of a poor quality array. The last quality measure is the 3'/5' ratios of the actin and graph. Affymetrix suggests that ratios below 3 show acceptable RNA degradation and recommend caution if that value is exceeded for a given array. There does not seem to be any problem with this data in terms of these statistics.

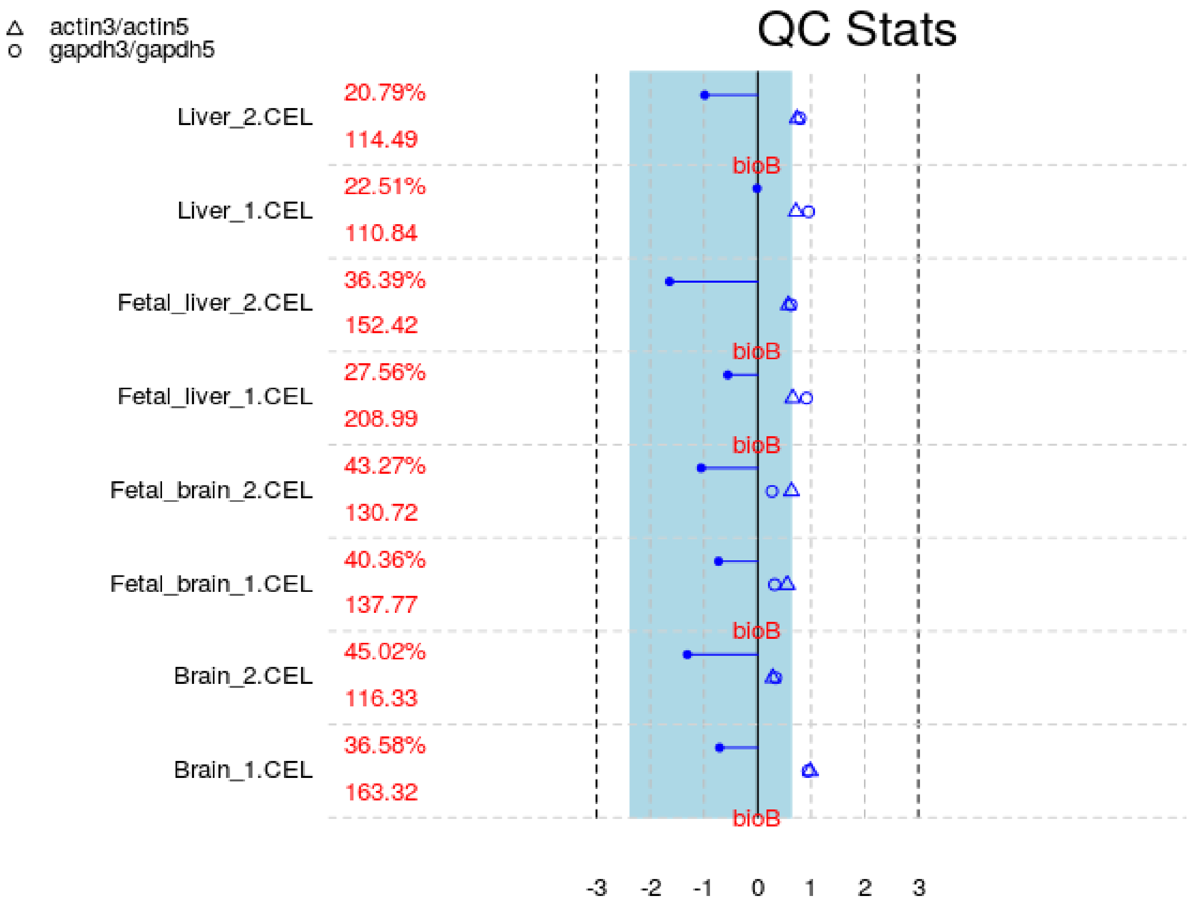

Figure 1.5: QC Statistics of raw data

Now let’s check the first five perfect match and mismatch intensities and their differences for the array brain 1 by using pm() and mm() functions. As you can see in *Figure 1.6*, sometimes the difference between perfect and mismatch values might result in negative intensities, which is impossible. Therefore, it is not a good idea to simply take differences of these two values as true intensity while doing background normalization.

	1	2	3	4	5
PM intensity	2122	1139	560	1739	644
MM intensity	3652	770	318	428	404
difference	-1530	369	242	1311	240

Figure 1.6: PM, MM and difference values for the first 5 observations of array Brain 1

Now we can start doing normalization on our data, we will use the rma() function. It consists of background normalization, log 2 transformation and quantile normalization steps. Now let’s check how the densities have changed. As you can see in *Figure 1.7* and *1.8*, all samples have similar probe densities now we have achieved quantile normalization successfully.

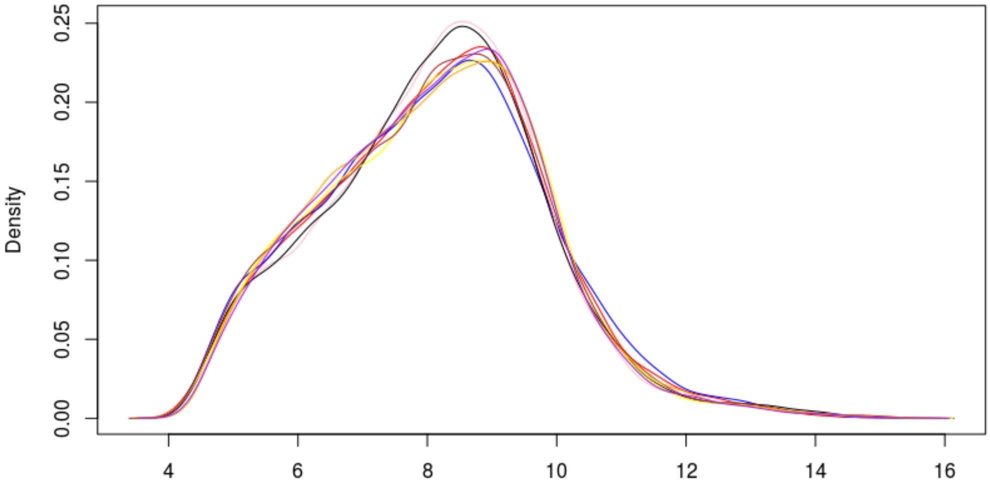

Figure1.7: Density plot of normalized data

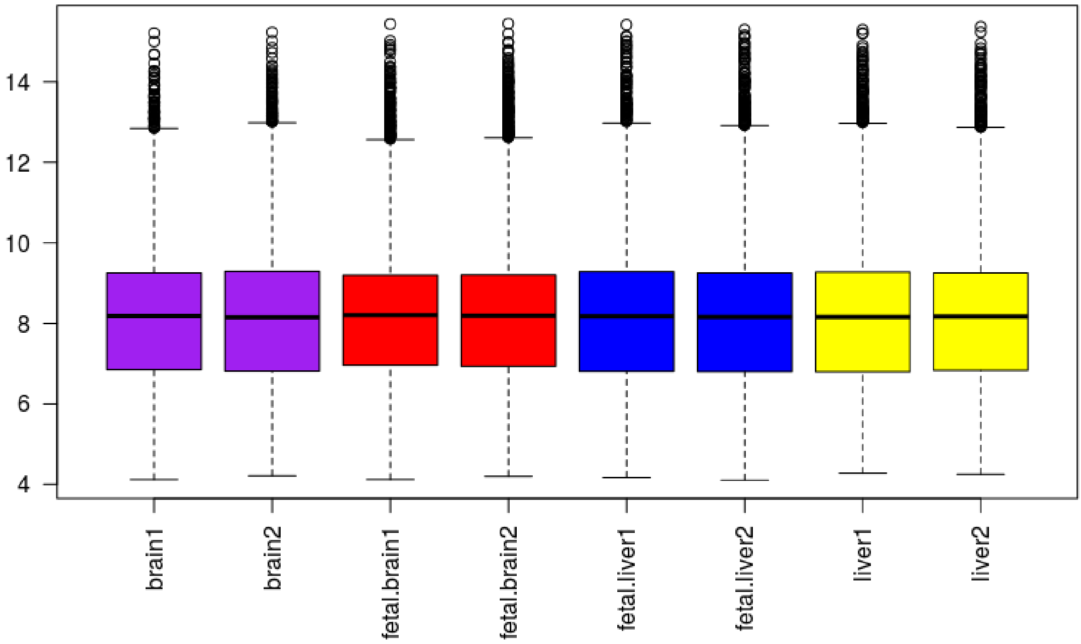

Figure1.8: Boxplot of normalized data

We can look at *Figure 1.9* for a comparison of normalized and raw data after only background normalization with log transformation. If there is no difference between raw and background corrected data, the data points should end up on the diagonal. We can see that effect on big intensities is not much compared to the background normalization effect on the small intensities.

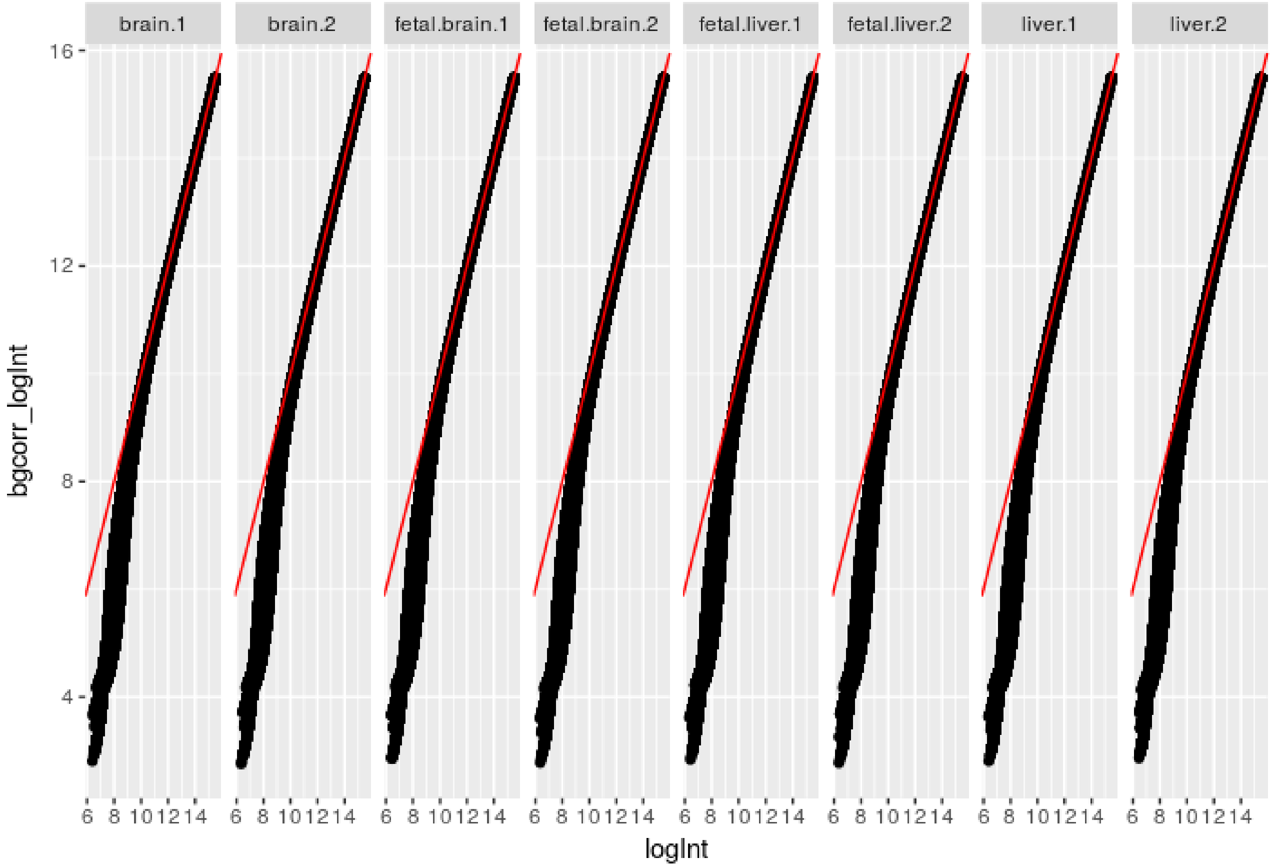

*Figure 1***Error! No text of specified style in document.***.9: Comparison plot for raw and background normalized data*

5)QUALITY CONTROL

After normalization, we need to check whether we achieved our goals in normalization processes. To test the quality, we should test the internal consistency of the data. We will try to assess the dissimilarity between arrays and visualize.

Principal Component Analysis: PCA is a dimension reduction method. We try to represent arrays in small dimensional space so that their similarities become easier to see. In PCA, we transform original dimensions to an orthogonal set which spans the same space, each of these orthogonal axes is called principal component. Among these principal components, the first principal component is the axis through the data along which there is the greatest variation amongst the objects. The second principal component is the axis orthogonal to the first that has the greatest variation in the data associated with it and so on. By using only first or first two principal components one can see the groups although there is some loss of information. If the arrays with same conditions are far from each other and do not form a group we can suspect that there is an internal inconsistency in our data. As you can see in *Figure 1.10*, when we apply PCA on our data samples with same conditions are making groups.

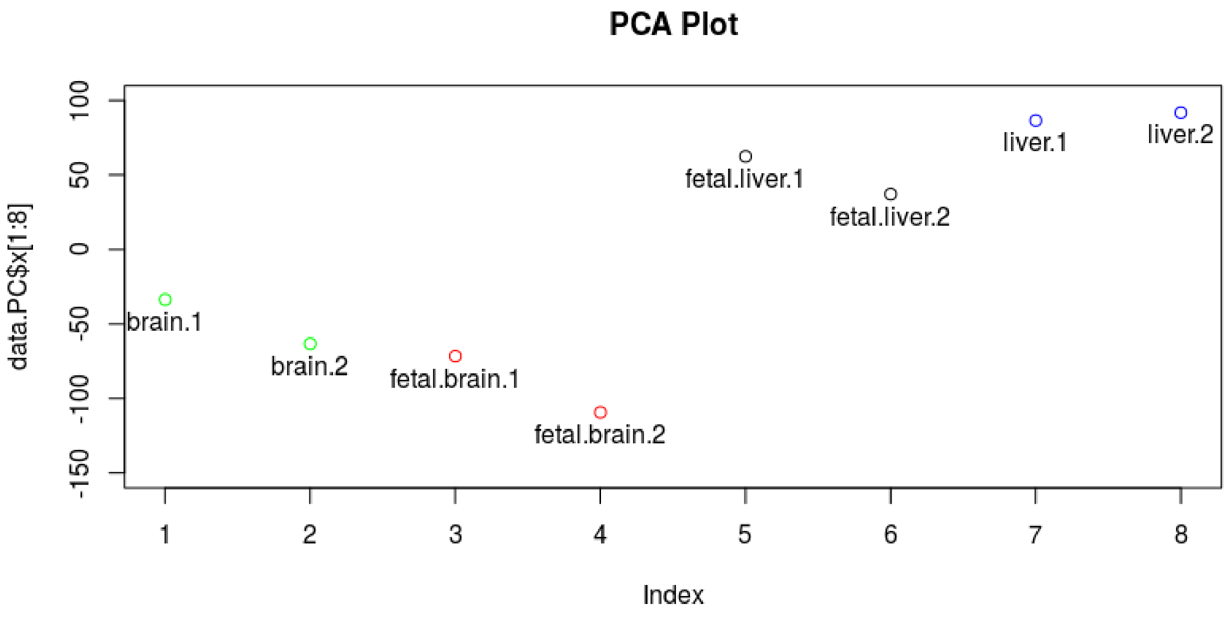

Figure 1.10: PCA plot

Pairwise Scatter Plots of Replicates: When plotting the normalized intensities via multiple pairwise scatter plots(MAs), points should scatter around the line of equality of the arrays. Deviations from this line indicate problems with internal consistency. For example, when we plot MVA plot of brain 1 and brain 2 after normalization, points are around the line of equality unlike the unnormalized case which is shown in *Figure 1.4*.

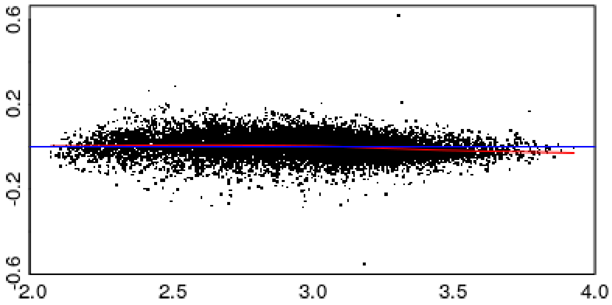

Figure 1.11: MVA plot of normalized brain sample 1 and brain sample 2

6)HYPOTHESIS TESTING

When we try to find differentially expressed genes, some hypothesis testing has to be applied with the null hypothesis: Gene is differentially expressed, i.e, active. In this project, like many other microarray analyses, we will be interested in differences in gene expressions between groups. So our null hypothesis will be the mean differences of two groups are statistically insignificant. Since a number of observations are very large we can use t-statistics to compare the means of two groups. We will apply t.test() function for comparing two groups, it performs Welch's test, so it is not assuming an equal variance between samples. Although after normalization process, as you can see in *Figure 1.8*, variances are almost same between samples, we use Welch test to be on the safe side. The p-values obtained for each gene, which is the raw values and they need to be corrected for multiple testing.

There are two different error rates in multiple testing that we will discuss. The familywise error rate (FWER) is the probability that among all inactive genes at least one is incorrectly classified as active. FWER is a very conservative error rate because especially when the number of hypotheses to be checked, denoted as n, is large, it is highly likely to make at least one incorrect classification of active genes. Among the methods we are going to apply Bonferroni and Hochberg procedures uses this rate. Because of that, they give a very small amount of active genes while trying to avoid even one incorrect classification. Another error rate, false discovery rate(FDR), is the expected number of inactive genes among those that are declared active. Unlike FWER, FDR allows us to tolerate a certain number of tests to be incorrectly classified. The third method that we are going to apply, Benjamini and Hochberg procedure, uses FDR instead of FWER.

Bonferroni procedure rejects the null hypothesis if the p-value is less than α/n, which is a very small value if n is large. Therefore, it is almost impossible to find an active gene. Hochberg procedure is less conservative compared to Bonferroni, it declares i^{th} gene active if after sorting the data, a raw p-value of this gene is less than α/ n-i+1. Although it is slightly corrected, it is not enough to declare many active genes when n is large. Last procedure Benjamini-Hochberg declares i^{th} gene as active if it is less than α*i/n. Number of active genes are higher in this procedure.

Unfortunately, when I applied all these three procedures, I get almost all genes as differentially expressed. I couldn’t figure out where the problem is. Therefore, for the sake of completeness, I will use raw p-values and 0.025 threshold now.

7)CONCLUSION

I applied t test to four different set of groups. Among adult brain tissues versus fetal brain tissues, the number of differentially expressed genes is 854 with 0.025 threshold value. Among adult liver tissues versus fetal liver tissues the number of significant genes is 591. Among all fetal tissues versus adult tissues this number is 538 and finally, among brain tissues versus liver tissues the number of differentially expressed genes is 3828.

From these results, we can conclude that brain tissues versus liver tissue has the most genetic expression difference. Adult brain versus fetal brain number is higher than adult liver versus fetal liver number, therefore it can be suggested that brain genetically differentiates more than liver throughout the growth of a human being. The number for fetal versus adult is smaller than both of the earlier groups, which is surprising. I would expect it to be somewhat in between these two numbers. If there were other sample tissues like muscle, heart etc., these tissues might have less changes than liver and brain, and they could pull back the number to a smaller value but now I don’t understand the reason behind this result.

References

http://manuals.bioinformatics.ucr.edu/home/R_BioCondManual
http://learn.genetics.utah.edu/content/labs/microarray/
http://jura.wi.mit.edu/bio/education/bioinfo2007/arrays/index.html
http://wiki.bits.vib.be/index.php/Analyze_your_own_microarray_data_in_R/Bioconductor
Wit, Ernst. *Statistics for microarrays design analysis and inference*. UK: John Wiley & Sons, Ltd, 2004.