2008-01-08

Sample comparisons using microarrays: - Application of False Discovery Rate and quadratic logistic regression

Ruijuan Guo
Worcester Polytechnic Institute

Sample comparisons using microarrays:
- Application of False Discovery Rate and quadratic logistic regression

by

Ruijuan Guo

A Project Report

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Master of Science

in

Applied Statistics

January 2008

APPROVED:

_____

Dr. Ryung S. Kim, Advisor

_____

Dr. Bogdan M. Vernescu, Department Head

Abstract

In microarray analysis, people are interested in those features that have different characters in diseased samples compared to normal samples. The usual p-value method of selecting significant genes either gives too many false positives or cannot detect all the significant features. The False Discovery Rate (FDR) method controls false positives and at the same time selects significant features. We introduced Benjamini's method and Storey's method to control FDR, applied the two methods to human Meningioma data. We found that Benjamini's method is more conservative and that, after the number of the tests exceeds a threshold, increase in number of tests will lead to decrease in number of significant genes. In the second chapter, we investigate ways to search interesting gene expressions that cannot be detected by linear models as *t*-test or ANOVA. We propose a novel approach to use quadratic logistic regression to detect genes in Meningioma data that have non-linear relationship within phenotypes. By using quadratic logistic regression, we can find genes whose expression correlates to their phenotypes both linearly and quadratically. Whether these genes have clinical significant is a very interesting question, since these genes most likely be neglected by traditional linear approach.

Acknowledgments

I would like to give my deep grateful to my advisor, Prof. Ryung S. Kim, for his help and patience. His suggestion and guidance are invaluable. I would also like to thank Prof. Joseph D. Petruccelli, Prof. Balgobin Nandram and Prof. Jayson D. Wilbur, I learned a lot from them. The one and half years I spend in Worcester Polytechnic Institute are very memorable. Finally I would like to thank my husband Hui Wang, without his love and support I couldn’t obtain where I am now.

Contents

List of Figures

List of tables

Chapter 1. Microarray Analysis Using FDR

1.1 Introduction

In microarray analysis, people are interested in those features that have different characters in diseased samples compared to normal samples. In order to test if a specific feature is significant or not, we need to select an appropriate test statistic, decide the significant level of the test, and compute the corresponding test statistic. One popular way to decide the significant features is to compare the p-value with the significant level of the test (e.g. α = 5% or 1%). If the p-value of the test is greater than or equal to α, wesupp conclude that the feature is significant; otherwise it is not significant. However, the p-value method is not practical in microarray analysis. For example, suppose we need to test *n* features at the same time, with each feature we control the test at significant level α. Then the family-wise significance level would be at most $1-(1-\alpha)^n$, which is greater than level α. Alternatively, we can set very low significant level for each test to make sure the family-wise significant level is low, however the resulting false positive rate will be high. Bonferroni (1936) correction can be used to test *n* independent tests at the same time and control the overall significant level at

α. It states that if an experimenter is testing n independent hypotheses on a set of data, then the statistical significant level that should be used for each hypothesis is $1/n$ times what it would be if only one hypothesis were tested. The main problem for using this method is that when the number of the tests n becomes larger, the significant level for each test will become smaller; eventually all the features are declared to be non-significant for sufficiently large n. So the Bonferroni method is too conservative when the number of hypothesis tests is large. To control the false positives, Benjamini (1995) and Storey (2003) proposed approaches to measure the statistical significance in the genome-wide studies based on the concept of False Discovery Rate (FDR). These approaches offer a sensible balance between the number of true and false positives. The first objective of this chapter is to introduce these two methods to control FDR. The second objective of our project is to apply Benjamini's and Storey's methods to analyze real data and compare their results. In addition, we will investigate what type of effect data transformation has on FDR methods.

In Section 1.2, we introduced the background to our project. The definition of FDR, and different methods to control the FDR are given in Section 1.3. In Section 1.4, we apply the methods to Meningioma data. The

effects of filtering criterion on FDR method will be discussed in Section 1.5. Conclusions and future work will be given in Section 1.6.

1.2 Scientific problem and data description

Meningioma is a type of brain tumor. The data in this report were collected by the experimenters in University of Texas Southwestern to study the relationship between genes and Meningioma types. In this project, we analyzed three groups of the Meningioma: A (mildest), B and C (most sever). We try to find the differentially expressed genes among the three groups.

By obtaining several samples from each cell type, we need to find genes that are differentially expressed among group A, B and C, where group A being the mildest and group C is the most severe. Our goal of this chapter is to use FDR of Benjamini's method and Storey's method to detect the significant genes, and compare the results of the two methods. What should be the cutoff for gene filtering criterion if the coefficient of variation is used as our filtering criterion?

Table 8 shows the data structure of our study; we first filtered out noisy genes based on coefficient of variation and obtain nested data sets with

different number of genes. For each data set there are three groups with sample size 7(A), 7(B) and 9(C), and then we took the log-transformation.

1.3 FDR and Sensitivity

FDR can be defined as the expected proportion of false positives among the declared significant results, which can be expressed as:

$$FDR = E\left[\frac{F(T)}{S(T)}\right] \qquad (1)$$

Sensitivity of the test using the FDR method is defined as the expected proportion of declared significant genes among the true significant genes. The statistics of the FDR method can best be described using Table 1 where 10,000 genes are classified according to their true status and the test result. In this example, the FDR is B/(B+D) = 475/875 = 54%, the sensitivity is D/(C+D)=80%, and the false positive rate (type I error) B/(A+B) = 475/9500 = 5%, which means that we have a test with 95% specification and 80% sensitivity, but more than half of the ‘discovered’ genes are false positive. This shows that the standard control of significant level leads to a high rate of false discoveries even when the power of the test would be considered adequate for a single-gene study.

In this section, we discuss the two methods to control FDR. One is Benjamini's method (1995) and another is Storey's method (2003).

1.3.1 Benjamini's Method

Benjamini's method for controlling the FDR includes the following steps: (1) Consider hypothesis H_1, $H_2 \ldots H_m$ based on corresponding P-values P_1, $P_2 \ldots P_m$. (m is the total number of the hypothesis tests), (2) Order P-values $P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(m)}$ and let $H_{(i)}$ be the corresponding hypothesis to $P_{(i)}$, (3) Let K be the largest I for which $P_{(i)} \leq$ (i/m)a. Reject all $H_{(i)}$, for I = 1,…,k, (4) Under the assumption that the test statistics are independent, it can be proved that this procedure controls FDR at level a.

The Benjamini's method is a conservative method, when tests for true null hypotheses are independent, this procedure will ensure FDR $\leq$ a.

1.3.2 Storey's Method

Storey's method for controlling FDR is more specific than Benjamini's method. In Storey's method, instead of defining FDR, we estimate FDR and then use the estimated FDR to control the tests. Let's first define a threshold t ($0<t<1$), where we call all features significant whose P-

value is less than or equal to t. If there are *m* hypothesis tests and we denote corresponding p-values by P_1, P_2 … P_m, then

$$F(T) = \#\{\text{null } P_i \leq t;\ i = 1 \ldots m\} \quad (2)$$

$$S(T) = \#\{P_i \leq t;\ i=1 \ldots m\} \quad (3)$$

We then can define

$$FDR(t) = E\left[\frac{F(T)}{S(T)}\right] \quad (4)$$

Because we are considering many features, it can be approximated by

$$FDR(t) = E\left[\frac{F(T)}{S(T)}\right] \approx \frac{E[F(T)]}{E[S(T)]} \quad (5)$$

A simple estimate of E [S(T)] is the observed S(T); that is the number of observed P-values less than t. In estimating E [F(T)], recall that p-values corresponding to truly null hypotheses should be uniformly distributed. Thus the probability a null P-values less than or equal to t is simply t, so E [F(T)] = m_0*t (m_0 is the true null). Because the total number of truly null features is unknown it has to be estimated. Equivalently, one can estimate the proportion of features that are truly nulls, which we denote by $\pi_0 = \frac{m_0}{m}$.

It is hard to estimate π_0 without specifying the distribution of the truly alternative P-values. However, using the factor that *p*-values of true nulls are uniformly distributed, a reasonable estimate can be formed. Storey defines

the point λ of the histogram of all p-values such that the distribution of p-values greater than the point the plot becomes flat: it means that there are mostly null p-values in this region. The height of this flat portion actually gives a conservative estimate of the overall proportion of null p-values. This can be quantified with

$$\hat{\pi}_0(\lambda) = \frac{\#\{P_i > \lambda; i = 1,...,m\}}{m(1-\lambda)} \tag{6}$$

Once we obtain the estimate of π_0, it is easy to obtain the estimate of FDR. The formula for estimating FDR is:

$$FDR(t) = \frac{\hat{\pi}_0 mt}{S(T)} = \frac{\hat{\pi}_0 mt}{\#\{P_i \le t\}} \tag{7}$$

And the sensitivity of the test is:

$$Sentivity = \frac{S(T)(1-FDR(t))}{m-m_0} = \frac{\#\{P_i \le t\}\left(1-\frac{\hat{\pi}_0 mt}{\#\{P_i \le t\}}\right)}{m-m\hat{\pi}_0} \tag{8}$$

1.4 Data analysis

1.4.1 Benjamini's method of controlling FDR

First, we will use Benjamini's method to control FDR to select the significant genes. We first filtered out genes with low coefficient of variation and obtained nested data sets with different number of genes/tests. We calculated *p*-values by performing ANOVA to each gene; the significant genes for each data set are obtained by controlling FDR at 0.1 level. The null hypothesis here is that the gene is not differentially expressed among group A, B and C. Table 2 shows the relationship of the number of tests and the number of significant genes. As we expected, the more the number of hypothesis tests, the less the number of significant genes selected.

We also studied the relationship between FDR and the statistic cutoffs. Using all 46713 genes, we performed ANOVA to each gene and obtained the *p*-value and F-statistic for each gene. By controlling FDR at a certain level we can calculate the p-value cutoff; we then compare the *p*-value of each gene and the *p*-value cutoff to decide if the gene is significant or not. The F-statistic cutoff can be calculated based on the *p*-value cutoff. Repeating above steps for different FDR levels, we can obtain the F-statistic cutoff for different FDR levels. The black dotted line in Figure1 shows the

relationship between FDR and F-statistic in Benjamin's method. Just for information, table 3 lists the top 15 significant genes selected.

1.4.2 Storey's method of controlling FDR

In order to use Storey's method, we need to estimate π_0 first. According to Storey (2003), there is a tradeoff between bias and variance in choosing the λ to use in $\pi_0(\lambda)$. It should be the case that the bias of $\pi_0(\lambda)$ should be decreasing with the increasing λ, the bias being the smallest when λ close to 1. Therefore, the method we used here is to estimate $\pi_0 = \lim_{\lambda->1} \pi_0(\lambda)$. As showed in Figure 2 for a range of λ, we plot λ versus $\pi_0(\lambda)$ and fit a cubic spline $\hat{f}$ for the data points then estimate the value of $\pi_0(\lambda) = \hat{f}(1)$. After we obtain the value of π_0, we can estimate FDR and sensitivity of test for different cutoff values.

Using the π_0 value obtained above, we can obtain the estimation of FDR and sensitivity using formulas (7) and (8) for different statistical cutoffs. The red line in Figure 1 shows the relationship between the estimated FDR and the cutoff of F-statistics in Storey's method. The blue dashed line shows the estimated sensitivity.

In order to see the effect of filtering to Storey's method, we obtained different number of significant genes by controlling FDR at level of 0.1 for various nested data sets. The results are presented in Table 4.

1.4.3 Comparison of Benjamini's method and Storey's method

As shown in the previous section, both Benjamini's method and Storey's method can be used to select the significant genes by controlling FDR at certain level. Our goal in this section is to investigate the difference between Benjamini's method and Storey's method.

Figure 1 demonstrates that Benjamini's method is more conservative than Storey's method : this is what we expected because Benjamini's method controls the upper bound of FDR. Figure 5 is the ROC curve, a plot of FDR versus sensitivity, for the two methods. From the figure, we again see that Storey's method is more sensitive than Benjamini's method, because at the same level of FDR, Storey's method has higher sensitivity than Benjamini's method. More interesting comparison of two methods are demonstrated in next section.

1.5 Investigation of the effect of data filtering

One of the common filtering criteria used by researchers is to analyze only genes with level of the coefficient of variation higher than certain level. Table 2 and Table 4 show that for the same number of tests and FDR level, using different methods the number of selected genes is different. Figure 4 plots the results from Table 2 and Table 4. From the Figure 1, we see that after the number of hypothesis tests reached certain threshold, increasing the number of tests will lead to decrease of the number of significant genes. It again shows that Benjamini's method is more conservative than Storey's method.

In the ideal case, all filtered-out genes are noisy when the filtering criterion (coefficient of variation) is less than a certain level. We might expect that the number of selected genes have nothing to do with the number of tests as long as we use Storey's method; while the number of selected genes should decrease when the number of the tests increases with Benjamini's method. This means that Storey's method should be more stable once the number of tests reached certain threshold. However, Figure 4 shows different result from what we expected. In this section we will investigate what may cause the difference.

We investigated the effect of changing the order of filtering and transformation. In all above analysis, we first filtered out the genes by coefficient of variation and made log transformation before performing hypothesis tests; but this may throw out some significant genes at the log scale. Table 6 gives the analysis result for the new analysis after we changed the order of filtering and transformation. Comparing Table 3 and Table 6, we can see that changing the scale for filtering process can affect the analysis result. In addition, by comparing the sensitivity of the two methods we can find that the second filtering scheme (first log-transform and then filter) has higher sensitivity than the first one (filter in original scale and then transform in log scale for hypothesis tests). Figure 7 and Figure 8 are the plots of number of significant genes versus number of all genes tested and plot of ROC respectively for the data analysis using the second filtering scheme. Comparing the plot of Figure 5 and Figure 8, we notice that for the same level of FDR the sensitivity of the test using the second filtering scheme is higher than using the first filtering scheme. This shows that filtering scheme do affect the analysis result.

In conclusion, we found that many significant genes in log scale have very low coefficient of variation in original scale. Figure 9 gives the box-plot of such an example. It shows an example gene that have CV less than

0.05 in original scale are differentially expressed between the three groups in log scale. The coefficient of variation for this gene is 0.038 in original scale, but the box-plot shows that the gene is differentially expressed between the three groups.

1.6 Conclusion and discussion

As we expected, Storey's method is more sensitive than Benjamini's method, because the sensitivity of the test at the same level of FDR is higher using Storey's method. And when the number of hypothesis tests reaches certain threshold, increasing the number of hypothesis tests will lead to decreasing the number of significant genes.

In addition, the order of filtering and transformation can affect the analysis results. We need to be careful in what scale the computation of coefficient of variation is performed. In this project, we investigated to find that many significant genes in log scale have very low coefficient of variation in original scale. Therefore, we need to pay attention to this method to filter genes.

Proper scaling of expression indices from microarray is critical however not enough attention has been given to this aspect. Most widely recommended scaling is log transformation. However, change in expression

of genes with high expression levels may lead to different test result. The proper scaling may be different according to the mean expression level. In the future, we may study the effect of Box-Cox transformation and then find a better way to filtering the noise genes. Another remedy is to use estimating non-linear relationship between expression and sample labels. In the next chapter of this project, we will discuss using quadratic logistic regression to select genes.

Chapter 2. Application of logistic regression in microarray analysis

2.1 Introduction

In microarray analysis, usually several samples for each phenotype of a disease are given. We are interested in selecting genes that are differentially expressed between phenotypes. There are different methods to detect these genes. The most popular methods are to use multiple *t*-test or ANOVA to obtain the significant genes as we discussed in chapter 1. But such linear models, can only detect genes that have linear relationship between the response variable and predictors. However, expression of some biologically meaningful genes may have non-linear relationship with phenotypes; such genes cannot be detected using *t*-test. These genes can be detected using the quadratic logistic regression method, which we will discuss in this chapter. We will compare the quadratic logistic regression result with linear methods. In this chapter we again use the same Meningioma data from chapter 1, but we only consider group A (mild status) and group C (the most sever status).

2.2 Logistic regression models

Linear logistic model:

$$\ln\frac{\pi_i}{1-\pi_i} = \beta_0 + X_i\beta, i = 1,2,...,n.\text{independent}$$
$$Y_i \sim Bernuolli(\pi_i) \quad (9)$$

Quadratic logistic regression model:

$$\ln\frac{\pi_i}{1-\pi_i} = \beta_0 + X_i\beta_1 + X_i^2\beta_2, i = 1,2,...,n.independent$$
$$Y_i \sim Bernuolli(\pi_i) \quad (10)$$

In both the linear (equation 9) and quadratic (equation 10) models, $Y_i = 1$ if the sample is in group C and $Y_i = 0$ if the sample is in group A. X_i is the log-transformed expression index for each gene in sample i. We assume that the gene expressions from all the samples are independent from each other.

2.3 Analysis Results

We fitted each gene in the data set using linear and quadratic logistic regression. We decided if the quadratic logistic regression model is suitable for the gene based on the *p*-value (p_2) obtained from the corresponding ANOVA (i.e. H_o: $\beta_1 = \beta_2 = 0$). Later for internal investigation, we also tested

if the quadratic logistic regression is more appropriate than linear logistic regression based on the *p*-values (p1vs2) obtained from ANOVA (i.e. H_o: $\beta_2 = 0$). By controlling FDR at 0.05 level we can obtain the significant genes base on p_2. The expression of significant genes we obtained based on the quadratic logistic regression method may have non-linear relationship with the phenotypes. Many of these genes cannot be detected by the usual *t*-test method.

In order to compare t-test and quadratic logistic regression method, we applied *t*-test to each individual gene. First we obtained the *p*-values of multiple *t*-test of the sample means of their expression indices in group A and C. Based on these *p*-values, by controlling FDR at 0.05, we obtained 158 significant genes. We found that all the 158 significant genes are in the group of significant genes selected by quadratic logistic regression. In summary, among all the 46713 genes, 1395 significant genes are detected using quadratic logistic regression at FDR 0.05, 158 significant genes are detected using *t*-test at FDR 0.05. We also selected top 1000 genes with the smallest *p*-values from *t*-test and top 1000 genes with the smallest *p*-values from quadratic logistic regression. By comparing the 2000 genes, we found that there are 415 genes overlapped in both top 1000 genes. Table8 shows the relationship between *t*-test and quadratic logistic regression results. In

figure 11, I gave several examples plots of gene expressions that can be detected by quadratic logistic regression but cannot be detected by *t*-test. From the plot, we can clearly see that the curvature plot instead of linear plot better describes those genes selected by quadratic logistic regression.

Figure12 is a plot of one specific significant genes selected by quadratic logistic regression. We can clearly see samples with mid-range expression have lower chance of cancer. We drawn the box-plot of a significant gene selected by quadratic logistic regression to see if there is any difference between the sample means of the gene expression indices of group A and C. The box-plot shows that there is no difference between the means of the two groups, which means that this gene cannot be detected by t-test.

From the above results we can conclude that there are some genes that are nonlinearly correlated between group A and C, so cannot be detected by *t*-test. But they can be detected by quadratic logistic regression. And the quadratic logistic regression is an appropriate method to select genes, which have significant curvature relationship between response variable and predictors.

2.4 Conclusion

In microarray analysis, there are many genes that their expression may be non-linearly correlated with the phenotypes. These genes cannot be found by linear tests such as *t*-test or ANOVA. By using quadratic logistic regression, we can find genes whose expression correlates to their phenotypes both linearly and quadratically. Whether these genes have clinical significant is a very interesting question, since these genes most likely be neglected by traditional linear approach. Quadratic regression is an appropriate method to select genes, which have curvature relationship between response variable and predictors.

References

1. Kutner, Nachtsheim and Neter (2004) “Applied Linear Regression Models ” *Mc Graw Hill.*
2. John D. Storey and Robert Tibshirani (2003) “Statistical Significance for Genomewide Studies” *PNAS, volume 100, no.16.*
3. Benjamini, Y. and Hochberg, Y. (1995) “Controlling the false discovery rate – a practical and powerful approach to multiple testing” *J. Roy. Stat. Soc. B Met., 57 (1): 289 – 300.*
4. Bonferroni, C.E. (1936) “ Teoria Statistica Delle Classi e Calcolo Delle Probabilita” *Istituto Superiore di Scienze Economiche e Commerciali de Fireze 8, 3-62*.

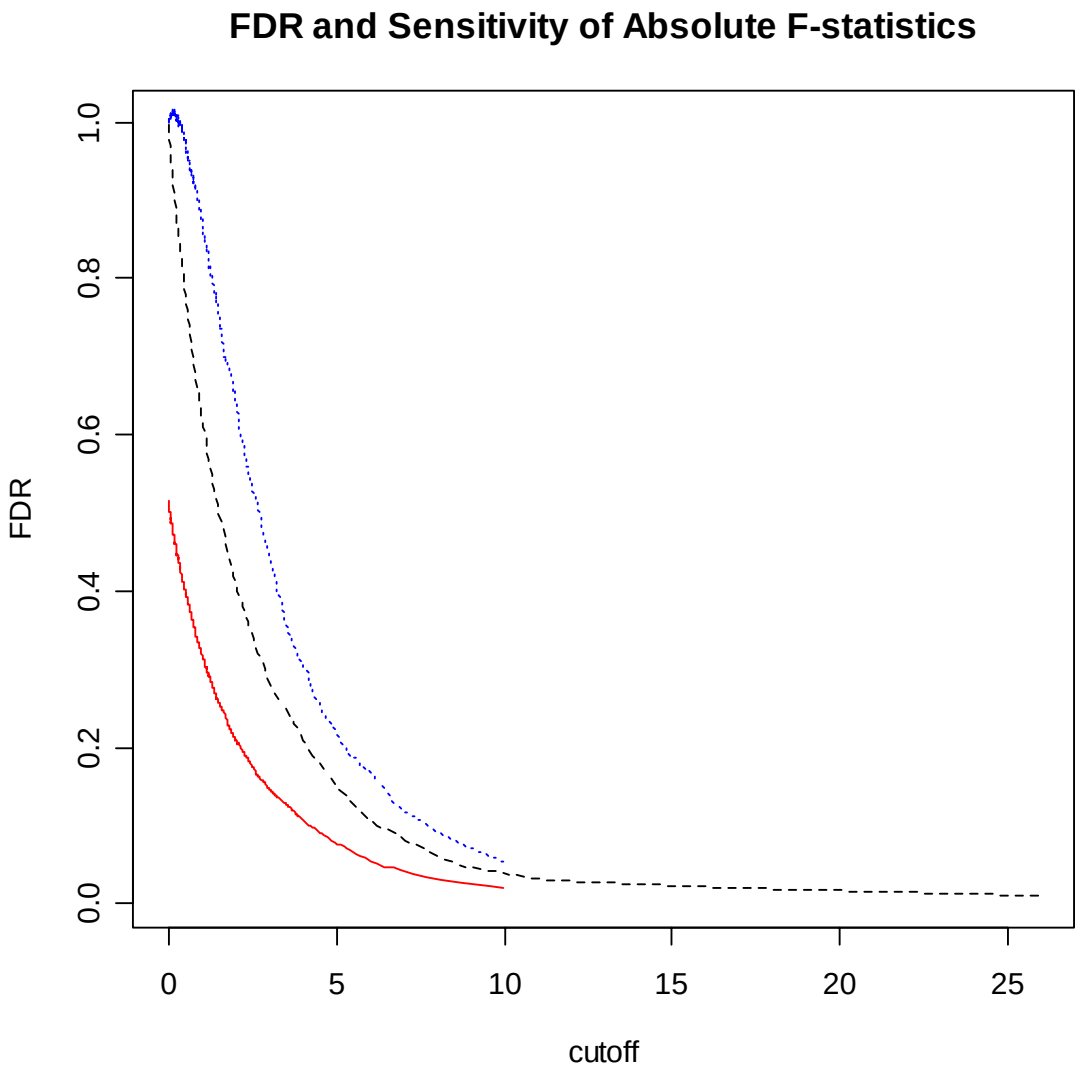

Figure 1, Plot of FDR and Sensitivity versus F-value cutoff. The red line is the curve of Storey's FDR versus cutoff; the blue dot line is the Storey's Sensitivity versus cutoff; the black slash line is the curve of Benjamini's FDR versus cutoff.

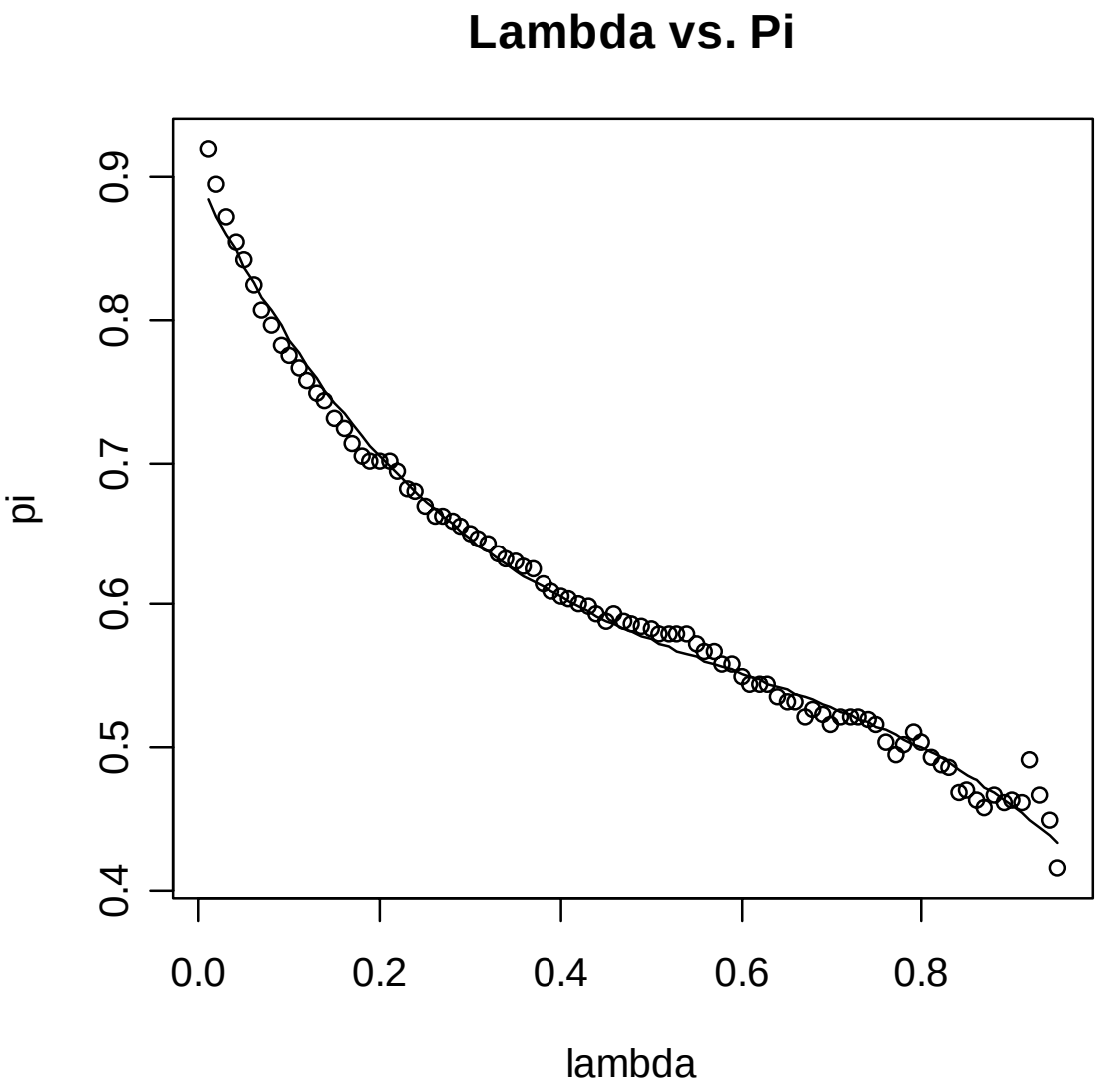

Figure 2, The $\pi_0(\lambda)$ versus λ for the data set with 46713 number of test.

The solid line is a cubic spline fit to these points to estimate $\pi_0(\lambda = 1)$

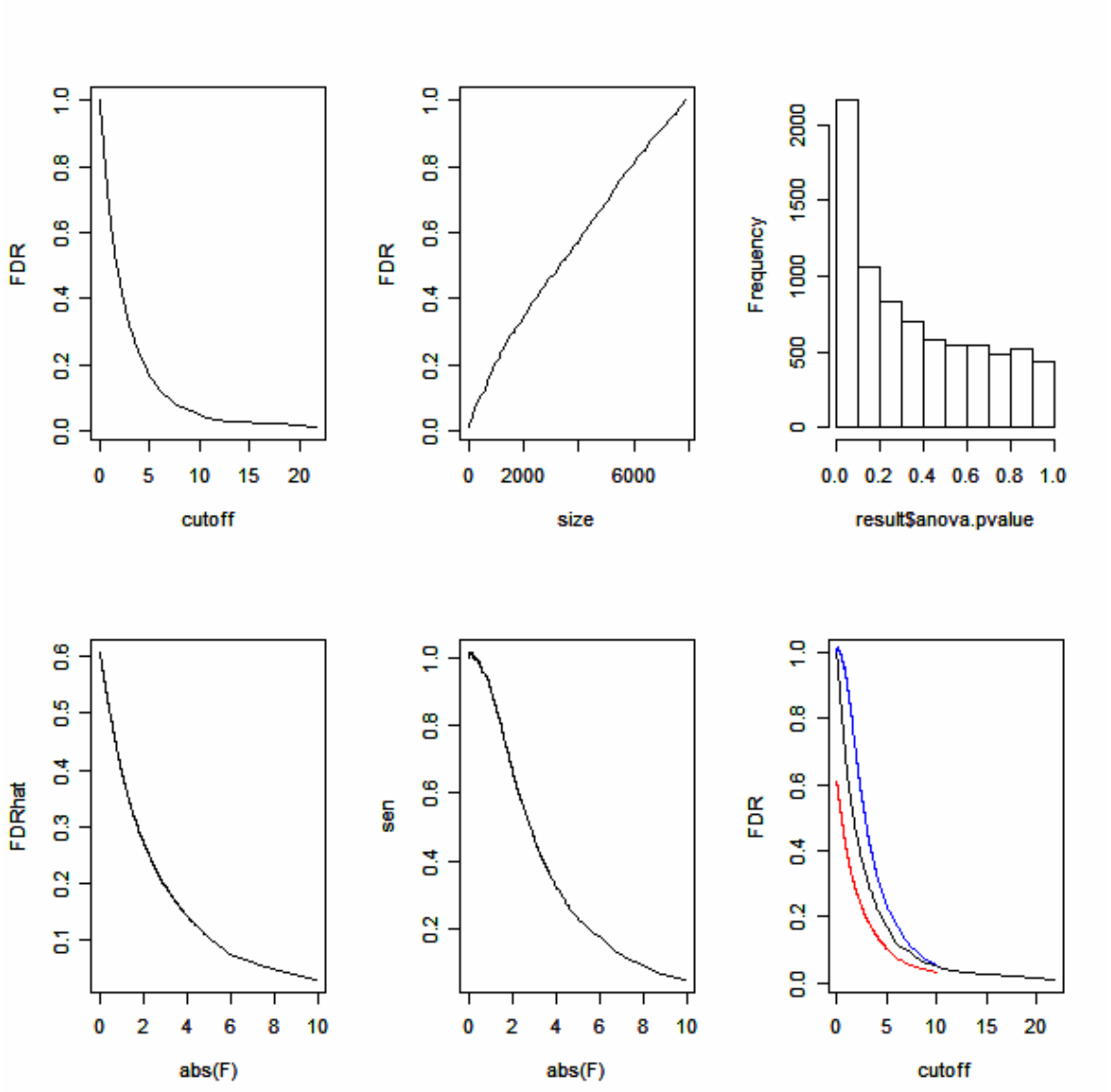

Figure 3, Top left is the plot of cutoff vs. FDR for Benjamini's method; top center is number of tests vs. FDR; top right is the histogram of the *p*-values; bottom left is cutoff vs. FDR for Storey's method; bottom center is cutoff vs. sensitivity for Storey's method; bottom right is cutoff vs. FDR and sensitivity for the two methods.

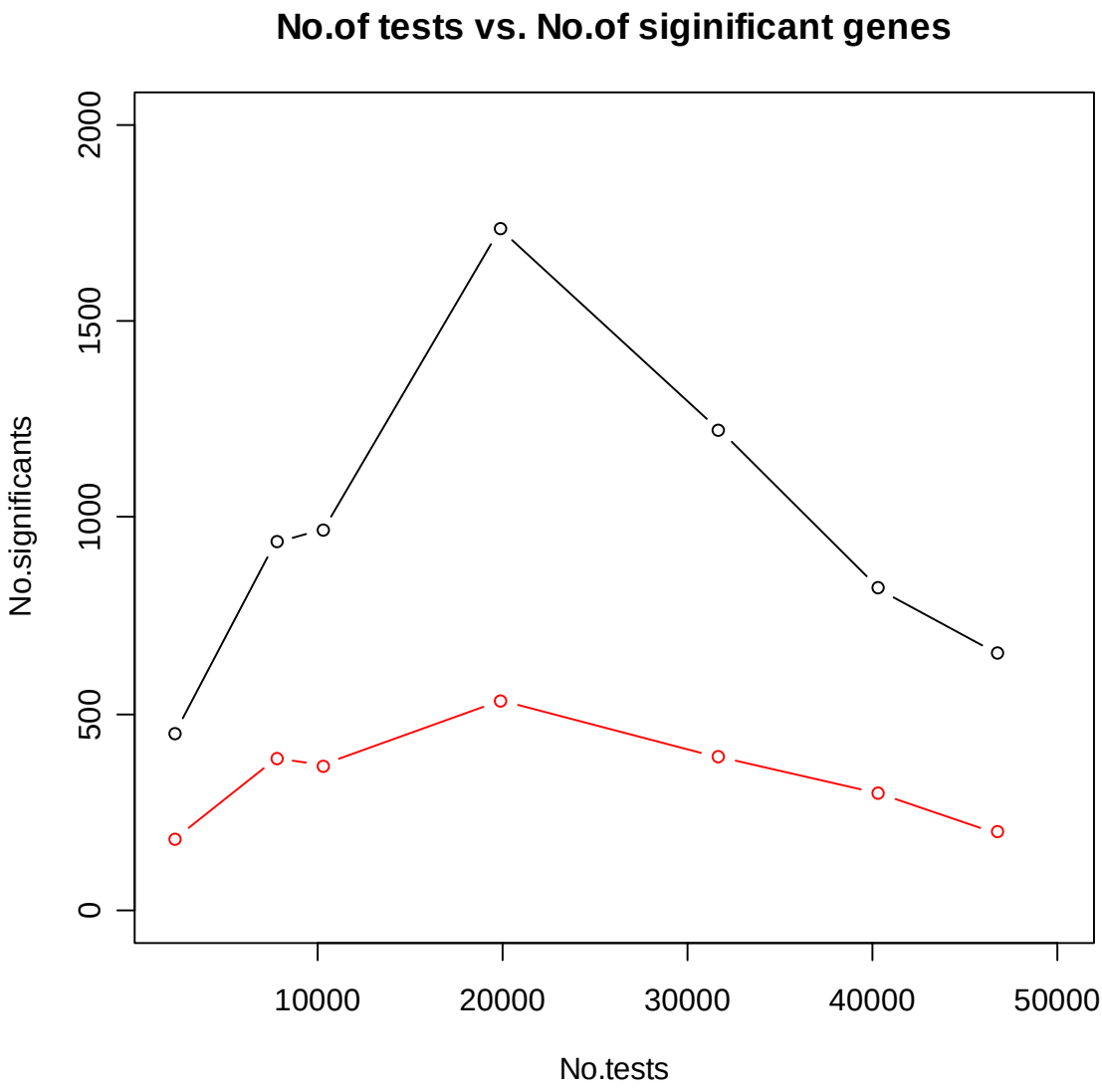

Figure 4, Here the genes are filtered first and then log transformed. Number of all genes tested versus the significant genes using two FDR methods controlling FDR at 0.01.The red line plot of number of test versus significant genes using Benjamini's method, the black line is the plot using Storey's method.

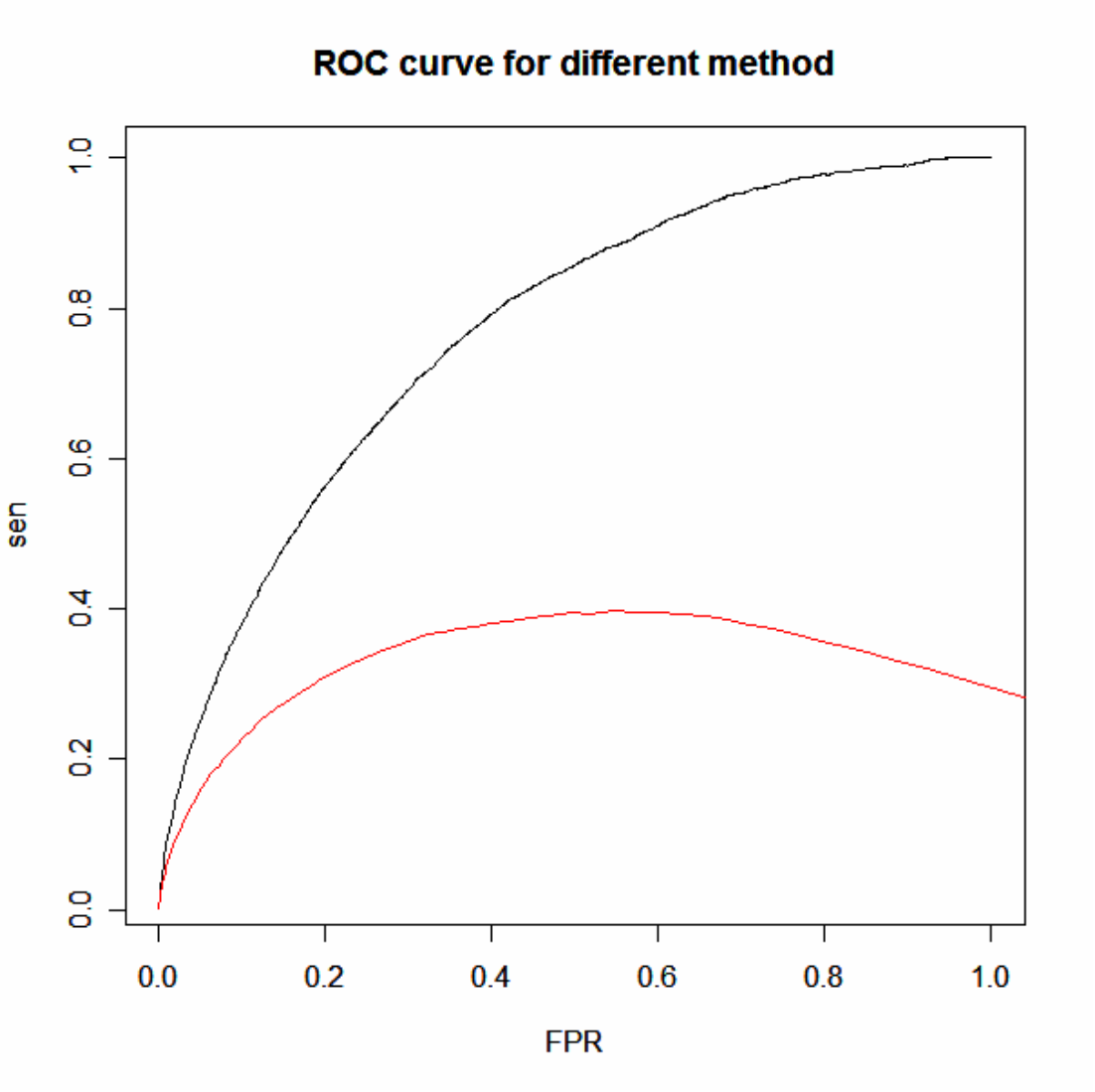

Figure 5, Here the genes are filtered first and then log transformed. Plot of FDR versus sensitivity using different method. The red line is plot of FDR versus sensitivity using Benjamini's method; the black line is the plot of FDR versus sensitivity using Storey's method.

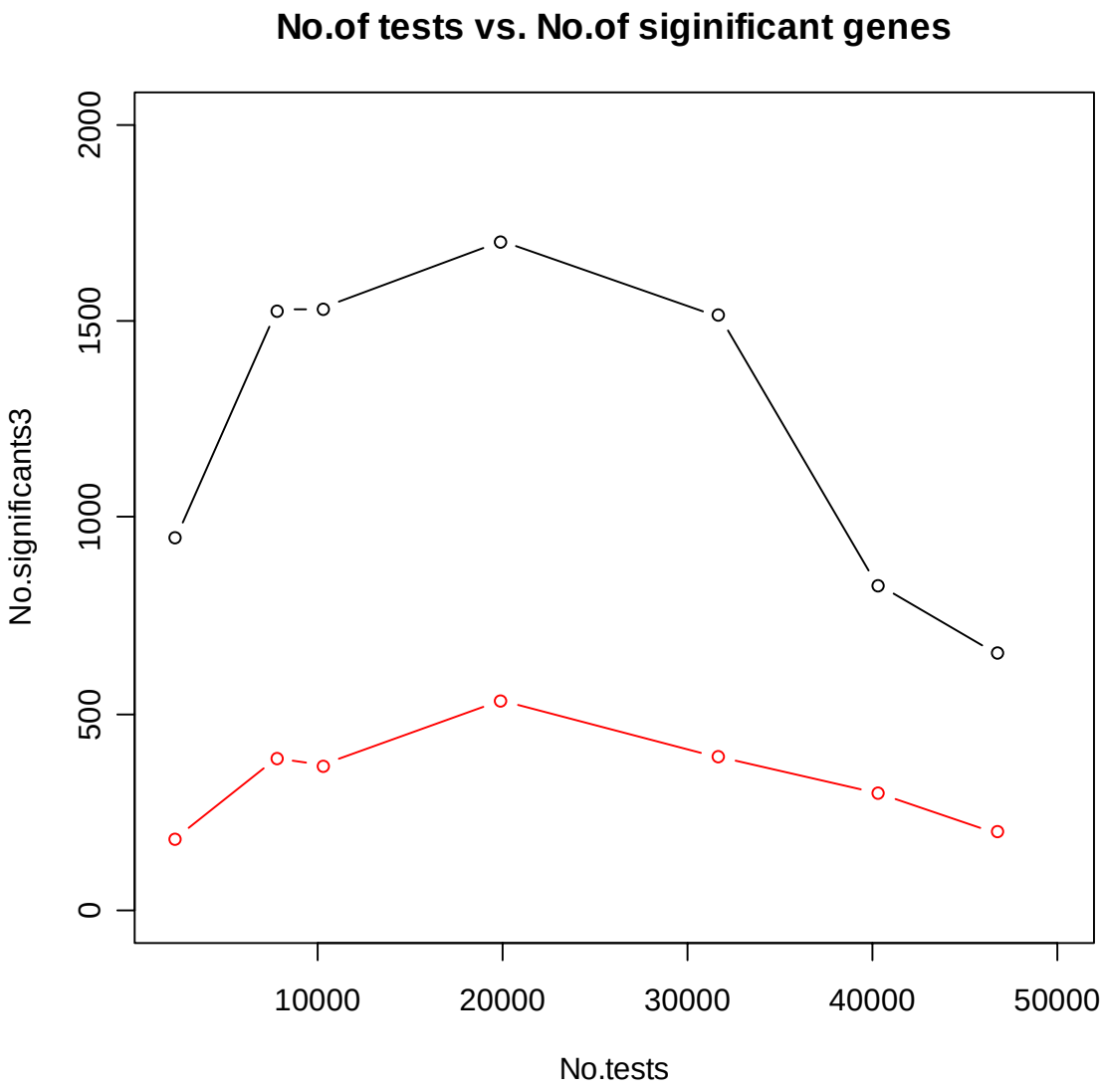

Figure 6, Here the genes are log transformed first and then filtered. Number of all genes tested versus the significant genes using two FDR methods controlling FDR at 0.01. The red line plot of number of test versus significant genes using Storey's method, the black line is the plot using Benjamini's method.

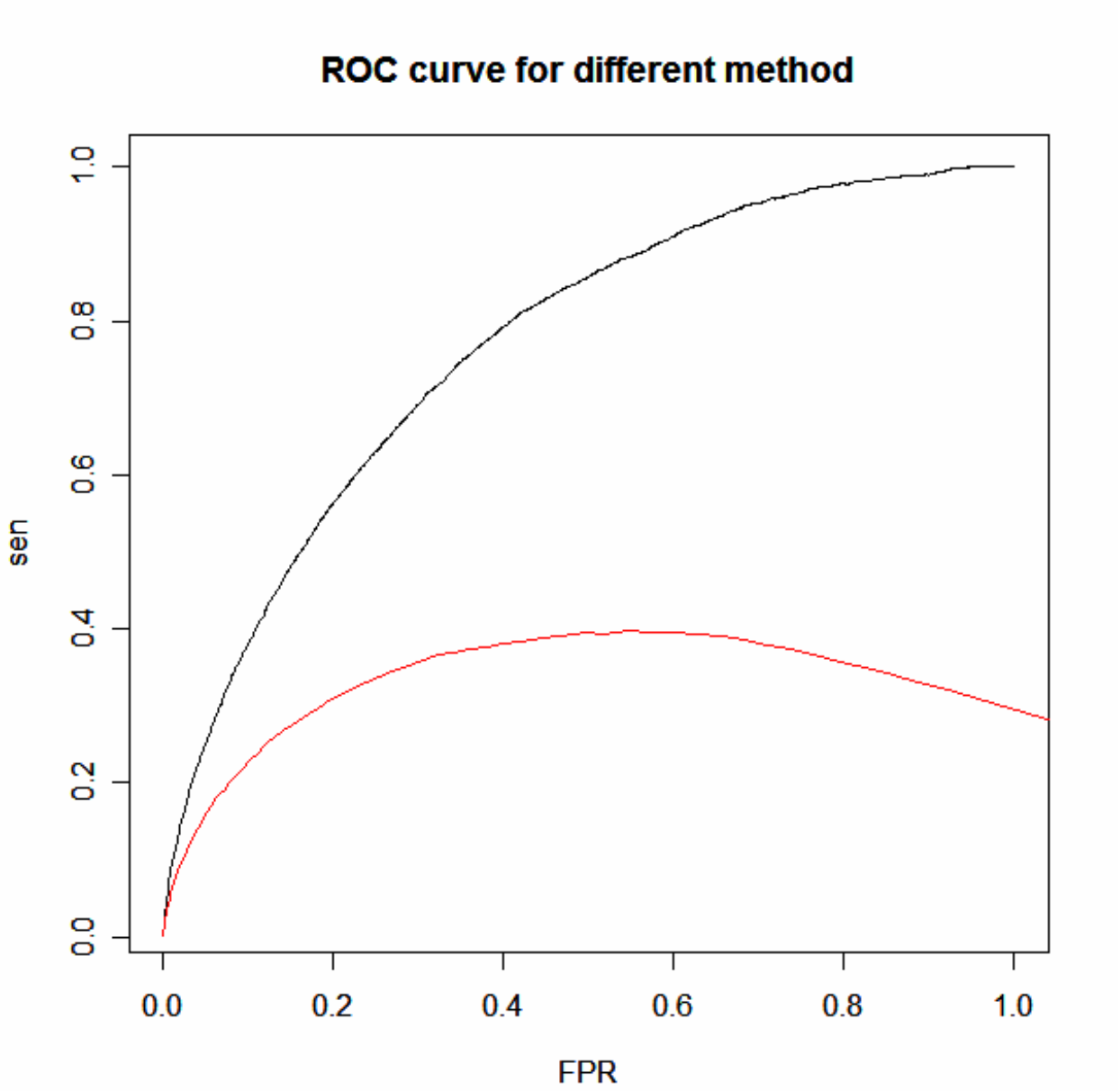

Figure 7, Here the genes are log transformed and then filtered. Plot of FDR versus sensitivity using different method. The red line is plot of FDR versus sensitivity using Benjamini's method; the black line is the plot of FDR versus sensitivity using Storey's method.

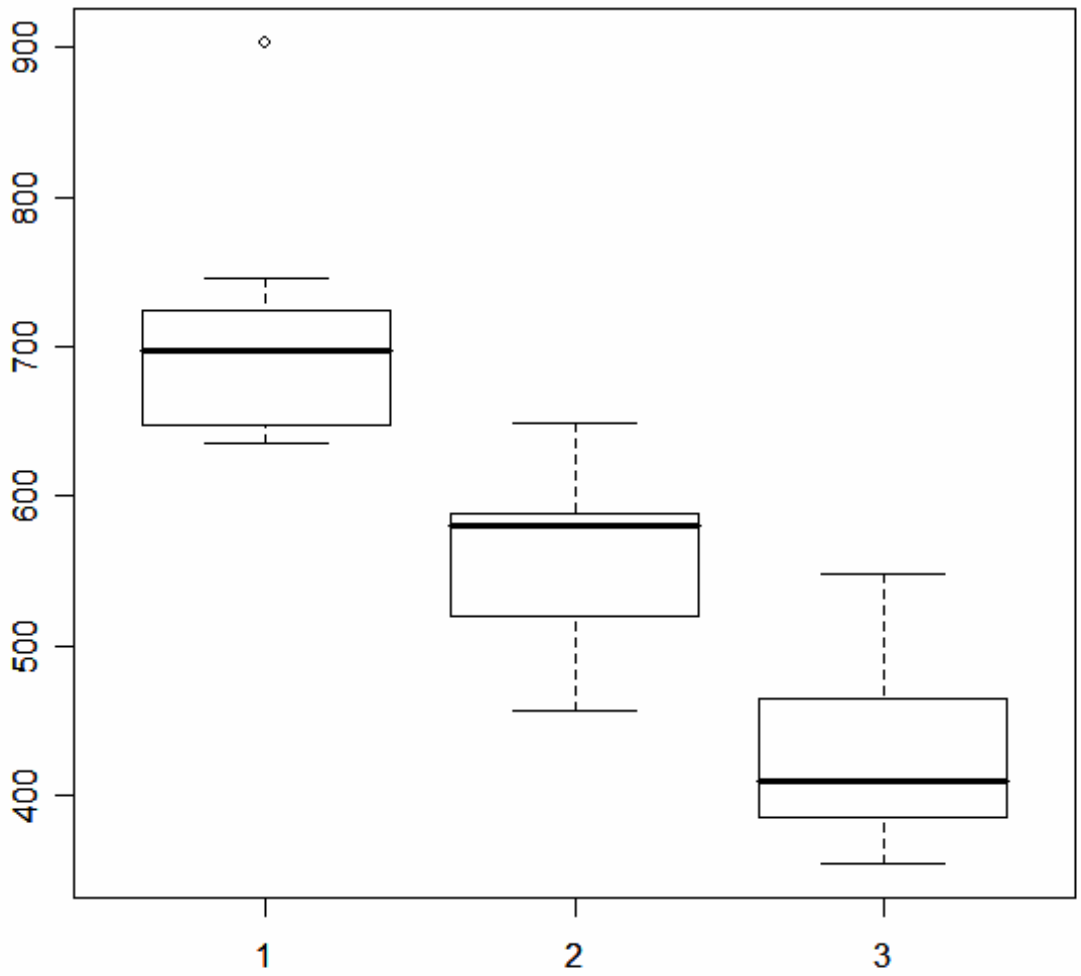

Figure 8, Box-plot of an example gene expression, which has small coefficient of variation in original scale 0.038 but large group difference in log scale.

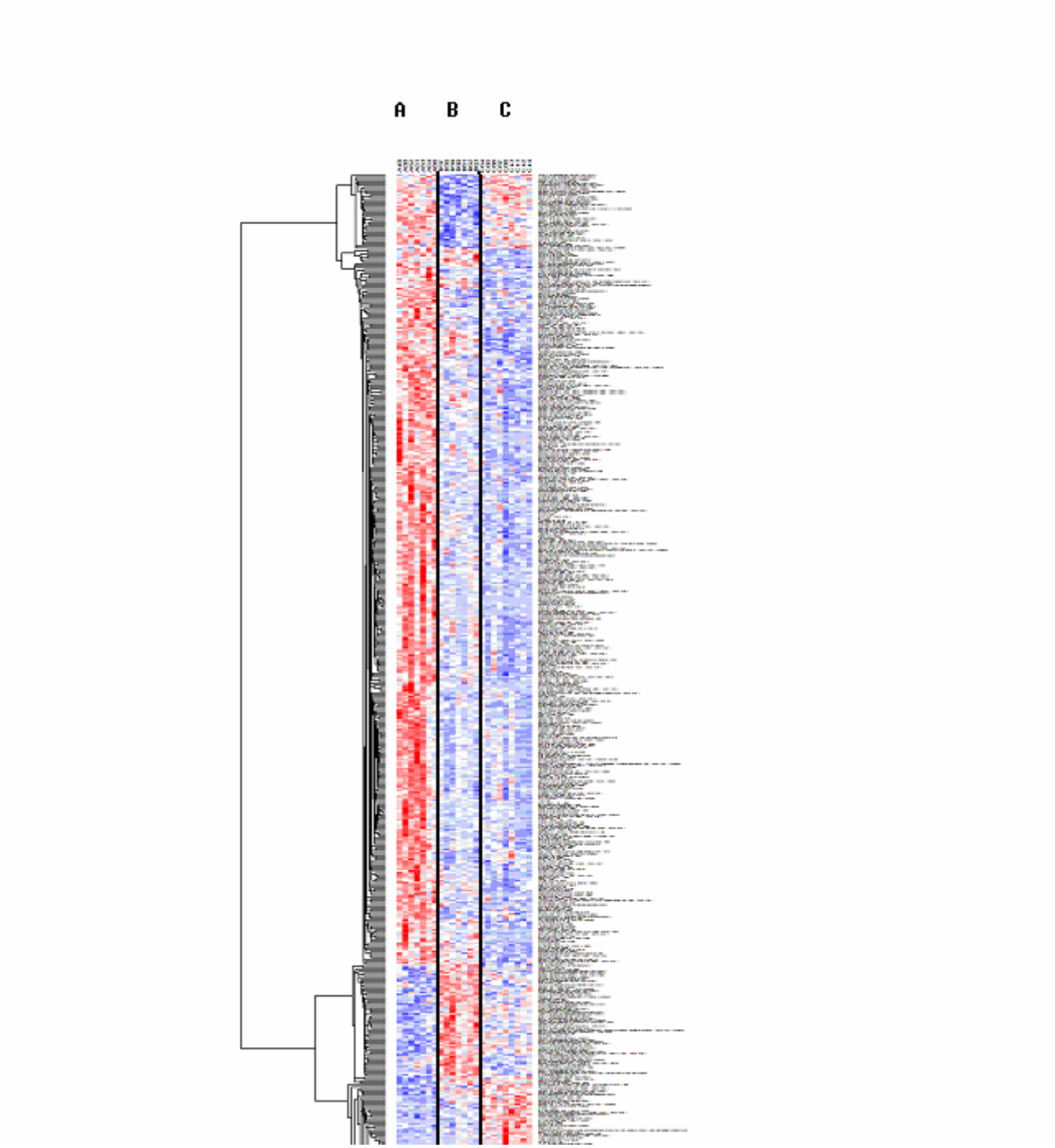

Figure 9, Heatmap diagram of clustering analysis result for the 655 significant genes in Table 4.

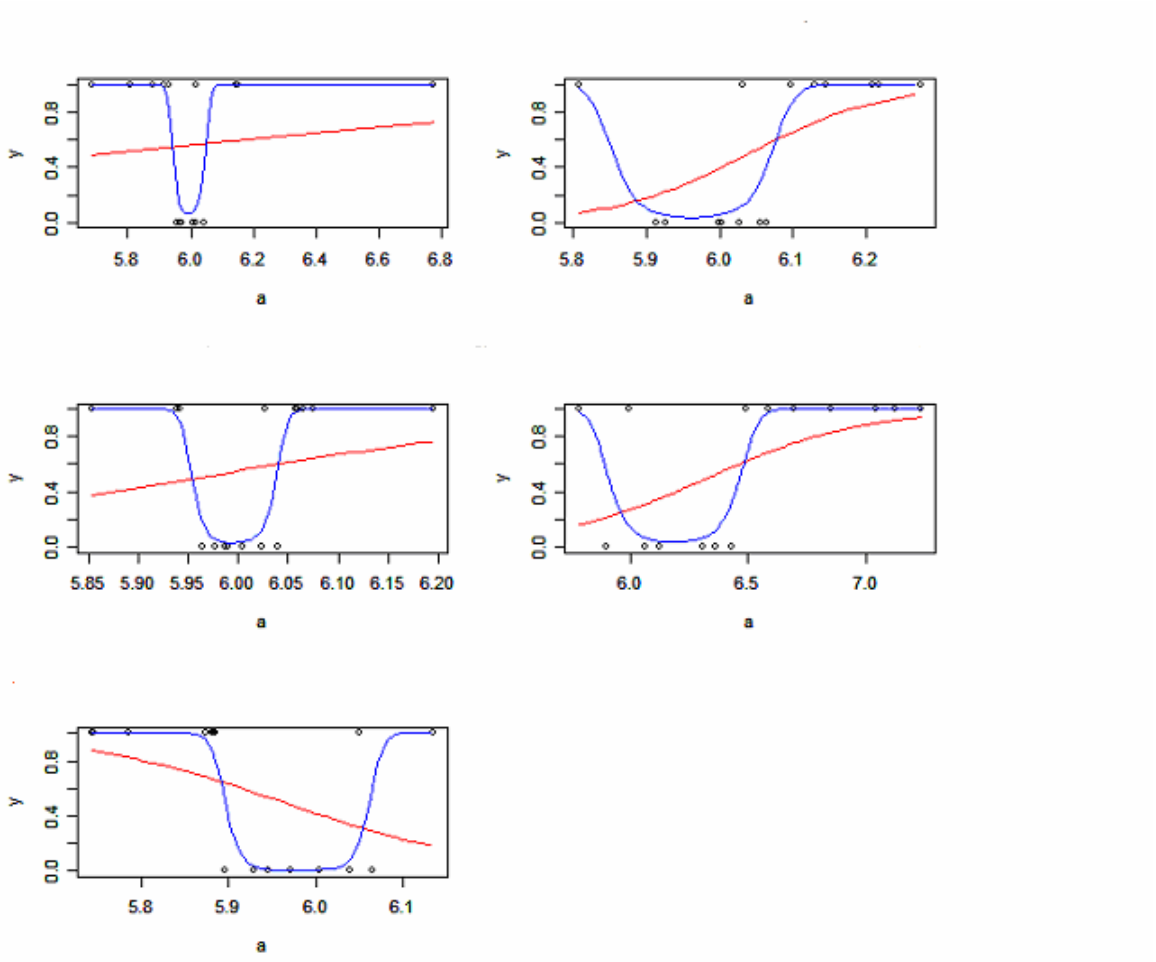

Figure 10, Plot of five example genes selected by quadratic regression method.

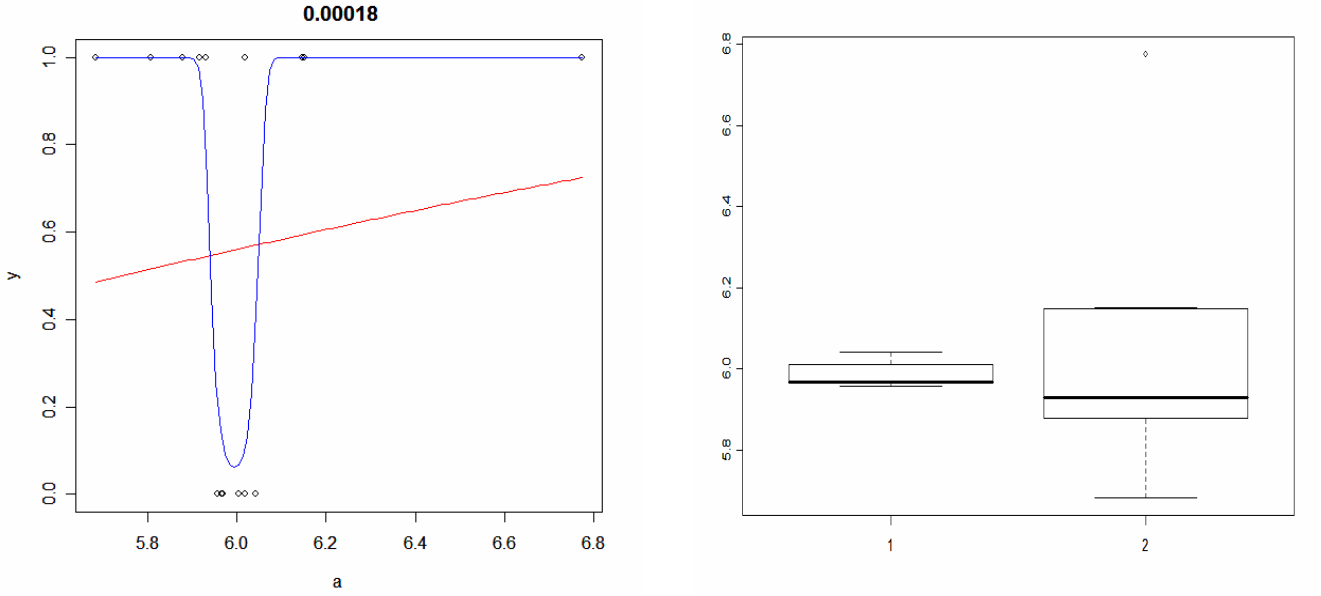

Figure 11, Plot of fitted regression model and box-plot for an example gene selected by logistic regression.

Table 1. A simple two-by-two table where 10,000 genes are classified according to their true status and the test result

	Test result		
	Non-DE	DE	Total
True			
Non-DE	A=9,025	B=475	9,500
DE	C=100	D=400	500
Total	9,125	875	10,000

Table 2. Number of significant genes selected for the same data set using different number of test when control Benjamini's FDR at 0.1 level.

Number of test	Significant genes
46713	200
40284	299
31640	392
19831	536
10246	370
7716	389
2264	184

Table3 gives the top 15 significant genes selected using Benjamini method:

Table3. Top 15 significant genes

	Probe.set	ANOVA.F	ANOVA.P	Q
37135	ILMN_5690	32.43616	5.28E-07	4.42E-05
582	ILMN_10261	28.53456	1.39E-06	8.83E-05
31629	ILMN_36696	25.89491	2.82E-06	1.33E-04
9502	ILMN_121343	19.12381	2.28E-05	1.77E-04
10205	ILMN_12337	18.44667	2.88E-05	2.21E-04
167	ILMN_10076	18.07336	3.29E-05	2.65E-04
36729	ILMN_5150	16.7922	5.25E-05	3.09E-04
42995	ILMN_8501	16.78053	5.27E-05	3.53E-04
27820	ILMN_29852	15.27435	9.40E-05	3.98E-04
22252	ILMN_22168	15.26481	9.44E-05	4.42E-04
24701	ILMN_25474	14.6905	1.19E-04	4.86E-04
20912	ILMN_20369	14.5146	1.28E-04	5.30E-04
18452	ILMN_17047	14.48072	1.29E-04	5.74E-04
42297	ILMN_8258	14.06279	1.54E-04	6.18E-04
10306	ILMN_12367	13.8236	1.70E-04	6.63E-04

Table 4. Related measurements for different number of test using Storey's method controlling FDR near to 0.1.

No. tests	π_0	M_0	$\overline{FDR}$	Sensitivity	Significant genes
46713	0.6286724	29368	0.1007958	0.03395507	655
40284	0.6098355	24567	0.1002720	0.04699743	821
31640	0.5613401	17761	0.1001688	0.07922603	1222
19831	0.4923403	9764	0.1030084	0.154497	1734
10246	0.5138269	5265	0.1010489	0.1743283	966
7716	0.4904709	3785	0.0999379	0.2154274	941
2264	0.402649	912	0.100306	0.3006957	452

Table 5. Top 15 significant genes selected using Storey's method

	Probe.set	ANOVA.F	ANOVA.P
37135	ILMN_5690	32.43616	5.28E-07
582	ILMN_10261	28.53456	1.39E-06
31629	ILMN_36696	25.89491	2.82E-06
9502	ILMN_121343	19.12381	2.28E-05
10205	ILMN_12337	18.44667	2.88E-05
167	ILMN_10076	18.07336	3.29E-05
36729	ILMN_5150	16.7922	5.25E-05
42995	ILMN_8501	16.78053	5.27E-05
27820	ILMN_29852	15.27435	9.40E-05
22252	ILMN_22168	15.26481	9.44E-05
24701	ILMN_25474	14.6905	1.19E-04
20912	ILMN_20369	14.5146	1.28E-04
18452	ILMN_17047	14.48072	1.29E-04
42297	ILMN_8258	14.06279	1.54E-04
10306	ILMN_12367	13.8236	1.70E-04

Table 6. Related measurements for different number of tests using Storey's method controlling FDR near to 0.1 after changing the order of filtering and transformation.

No. tests	π_0	M_0	$\overline{FDR}$	Sensitivity	Significant genes
46713	0.6286724	29368	0.100796	0.033955	655
39099	0.6027634	23568	0.099926	0.047925	827
22758	0.5203729	11843	0.100349	0.124702	1513
14970	0.4710356	7052	0.100089	0.193537	1703
8307	0.4319756	3589	0.099832	0.292261	1532
4081	0.2989897	1221	0.102538	0.478717	1526
1975	0.2598582	513	0.103219	0.582811	950

Table 7. Comparison of *t*-test result and logistic regression result.

	Significant genes		
	Logistic Regression	T-test	Overlap
FDR = 0.05	1395	158	158
Top 10000	1000	1000	415

Table 8. Data structure that is used

Sample Size			Number of Tests	Filter Criteria Variation of Coefficient
A	B	C		
7	7	9	2264	Larger than 0.5
7	7	9	7716	Larger than 0.25
7	7	9	10246	Larger than 0.1
7	7	9	19831	Larger than 0.09
7	7	9	31640	Larger than 0.07
7	7	9	40284	Larger than 0.06
7	7	9	46713	No