Tutorial Assignment 4

- Student Name: Zheyuan Zhou 周喆媛
- Student ID: 117010423
- Date: 2020.4.9

(a) Import Libraries

```
#!/usr/bin/env Python3
# coding=utf-8
from math import log
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib
import matplotlib.pyplot as plt
from scipy import stats
```

(b) Dataset Input

- Read the data from local file as excel: smokers.xls
- Rmoker is the dataset we apply analysis on, which for convenience only contain the ID_REF as primary key and drop the columns: Species Scientific Name and Gene Label
- Smoker_cp is the copy containing all the info of original file, in the case user want to find corresponding gene label

```
smoker_cp = pd.read_excel("./smokers.xls", sep="\t", index_col=0)
smoker = pd.read_excel("./smokers.xls", sep="\t", index_col=0)

smoker = smoker.drop("Gene Label", axis = 1)
smoker = smoker.drop("Species Scientific Name", axis =1)

print(smoker.head())
print("Data contains %d normalized probsets, %d samples." % smoker.shape)
```

```
            GSM101095    GSM101096    GSM101097    GSM101098    GSM101099  \
ID_REF
1007_s_at  3884.318400  1657.214200  2237.643600  1474.739300  2231.866000
1053_at      82.294170    74.921800    76.623764    54.349518    72.081345
117_at       37.470535    77.169880    27.224297    29.231043    30.802940
121_at      254.769700   173.070400   177.904650   135.697620   228.584460
1255_g_at     9.972142     9.346519    11.320443     8.536531    10.041258
```

```
             GSM101100    GSM101101    GSM101102    GSM101103    GSM101104  \
ID_REF
1007_s_at  2535.999500  1956.101300  2298.428000  2410.525400  2495.134300
1053_at      78.715040   101.283134   110.999010   114.482040    80.897100
117_at       26.526035    24.233334    24.874979    26.564205    32.034077
121_at      151.093810   168.721050   170.702000   149.034820   209.503080
1255_g_at     9.035363     9.285970    10.272593    10.256649     9.776113

            ...    GSM101107    GSM101108    GSM101109    GSM101110  \
ID_REF      ...
1007_s_at   ...  1862.330400  1844.315700  2097.944000  2466.514200
1053_at     ...    65.823784    58.283974    80.165850    89.279990
117_at      ...    39.444940    51.292150    32.644653    43.022903
121_at      ...   229.414550   245.255550   156.447740   168.562560
1255_g_at   ...     9.948459     9.965394    10.015328    11.016840

             GSM101111    GSM101112    GSM101113    GSM101114    GSM101115  \
ID_REF
1007_s_at  1992.740700  2232.849400  3326.898000  2238.369400  2615.229700
1053_at      83.198265    60.594906    76.577590    76.259480    68.664340
117_at       33.190834    26.173641    28.923296    26.477106    23.935286
121_at      135.459460   177.648970   190.692440   188.691470   171.054470
1255_g_at     9.567991    10.641581     9.388623     8.894799     9.506586

             GSM101116
ID_REF
1007_s_at  2130.132600
1053_at      60.522540
117_at       27.868433
121_at      129.480180
1255_g_at     9.235668

[5 rows x 22 columns]
Data contains 54613 normalized probsets, 22 samples.
```

(c) Data Reorganization

- Observe the data, it is easy to notice that the number magnitude is diversed: from thousand to unit, which is not good for further analysis and visualization.
- Apply Log 2 transformation to reorder the number magnitude.
- **Print the smoker dataset head again, notice the data has already been transformed.**

```
dataset =
["GSM101095","GSM101096","GSM101097","GSM101098","GSM101099","GSM101100","GSM1
01101",

"GSM101102","GSM101103","GSM101104","GSM101105","GSM101106","GSM101107","GSM10
1108",

"GSM101109","GSM101110","GSM101111","GSM101112","GSM101113","GSM101114","GSM10
1115",
           "GSM101116"]
for target_col in dataset:
    smoker[target_col] = np.log2(smoker[target_col])
print(smoker.head())
```

```
           GSM101095  GSM101096  GSM101097  GSM101098  GSM101099  GSM101100  \
ID_REF
1007_s_at  11.923446  10.694544  11.127765  10.526244  11.124035  11.308339
1053_at     6.362718   6.227314   6.259720   5.764195   6.171554   6.298567
117_at      5.227685   6.269966   4.766823   4.869429   4.944996   4.729337
121_at      7.993050   7.435215   7.474960   7.084252   7.836584   7.239301
1255_g_at   3.317903   3.224429   3.500859   3.093650   3.327868   3.175583

           GSM101101  GSM101102  GSM101103  GSM101104  ...  GSM101107  \
ID_REF                                                 ...
1007_s_at  10.933765  11.166432  11.235132  11.284902  ...  10.862893
1053_at     6.662250   6.794403   6.838977   6.338016  ...   6.040537
117_at      4.598921   4.636623   4.731412   5.001536  ...   5.301768
121_at      7.398496   7.415336   7.219506   7.710828  ...   7.841813
1255_g_at   3.215053   3.360728   3.358488   3.289261  ...   3.314473

           GSM101108  GSM101109  GSM101110  GSM101111  GSM101112  GSM101113  \
ID_REF
1007_s_at  10.848870  11.034760  11.268258  10.960538  11.124670  11.699962
1053_at     5.865027   6.324916   6.480265   6.378482   5.921125   6.258850
117_at      5.680666   5.028775   5.427033   5.052713   4.710043   4.854160
121_at      7.938142   7.289537   7.397140   7.081717   7.472886   7.575104
1255_g_at   3.316927   3.324138   3.461639   3.258216   3.411641   3.230914

           GSM101114  GSM101115  GSM101116
ID_REF
1007_s_at  11.128232  11.352722  11.056728
1053_at     6.252845   6.101489   5.919401
117_at      4.726674   4.581067   4.800560
121_at      7.559885   7.418312   7.016587
1255_g_at   3.152962   3.248927   3.207216

[5 rows x 22 columns]
```

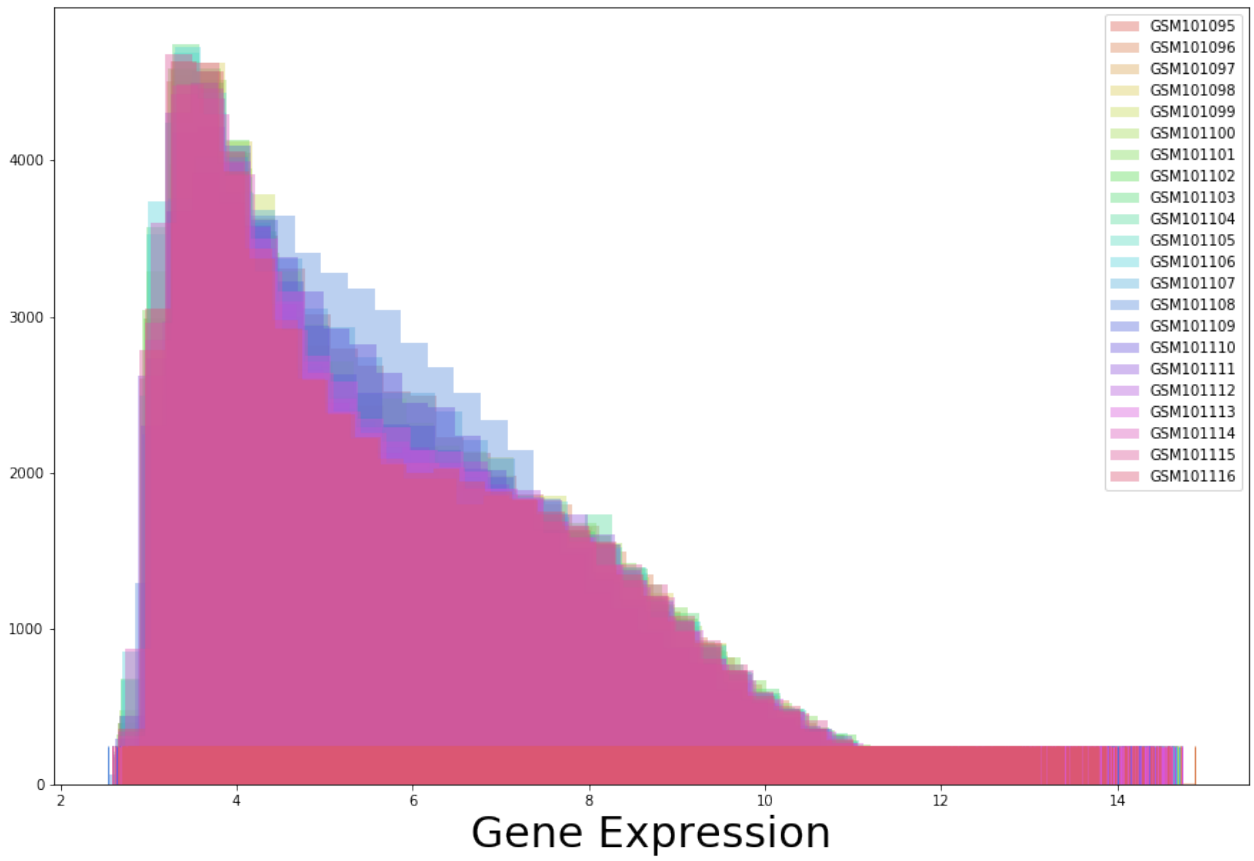
(d) Question 1

Plot the gene expression value distribution among 22 samples

- One Histogram Graph
- Separated Histogram Graph
- One Histogram Graph with Kernel Density Estimation
- Separated Histogram Graph with Kernel Density Estimation
- Box Plot

Plot a histogram of the subject group of 22 samples

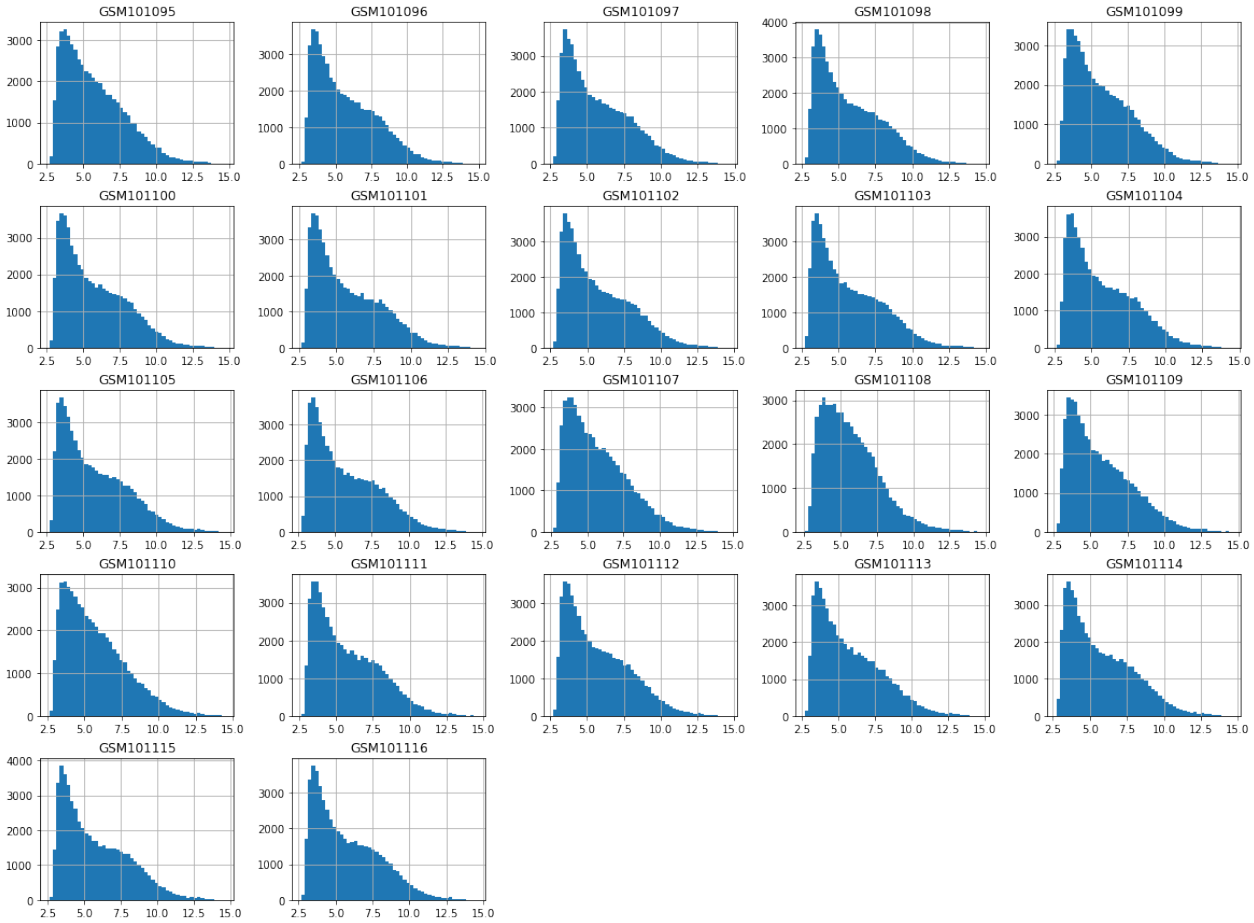
```
plt.figure(figsize=[15,10])
with sns.color_palette('hls',22):
    for target in dataset:
        ax = sns.distplot(smoker[target], bins=40, label=target, rug=True,
kde=False)
plt.xlabel("Gene Expression", size = 30)
lgd = plt.legend()
```



Plot a histogram for each sample respectively

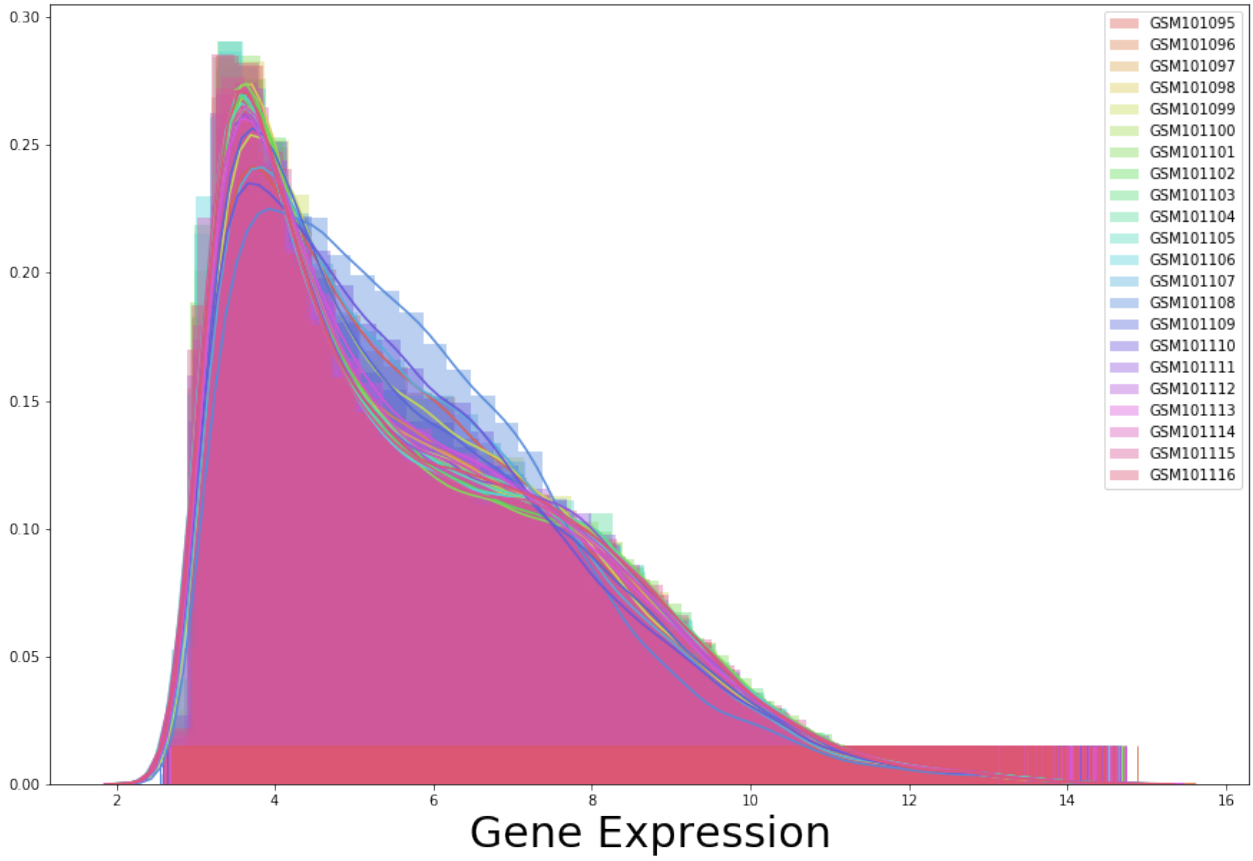
Notice the y axis range in these figures are almost the same, therefore it is still easy for us to compare the data.

```
smoker.hist(bins=50, figsize=(20,15))
plt.show()
```



Plot a histogram of the subject group of 22 samples with kde

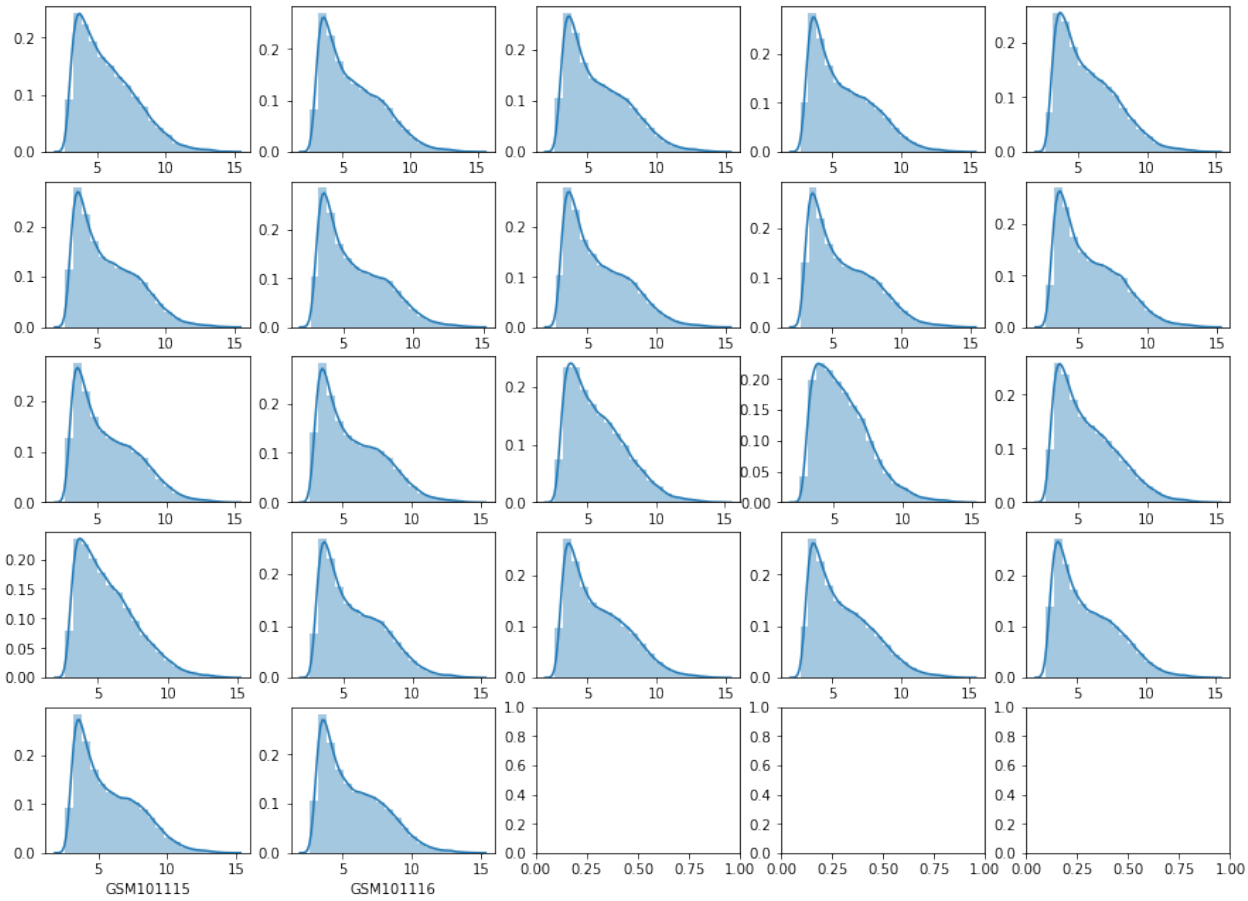
```
plt.figure(figsize=[15,10])
with sns.color_palette('hls',22):
    for target in dataset:
        ax = sns.distplot(smoker[target], bins=40, label=target, rug=True,
kde=True)
plt.xlabel("Gene Expression", size = 30)
lgd = plt.legend()
```



Plot a histogram with kde for each sample respectively

Notice the y axis range in these figures are almost the same, therefore it is still easy for us to compare the data

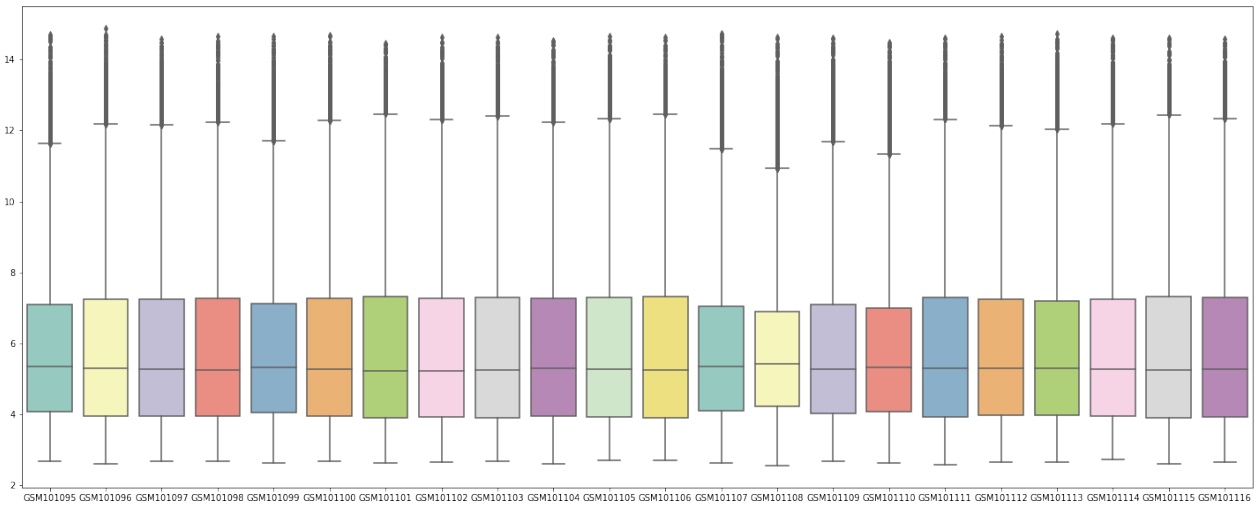
```
fig, axes = plt.subplots(5,5,figsize = (15,11))
count = 0
with sns.color_palette('hls',22):
    for target in dataset:
        ax.set_ylim(0,0.3)
        sns.distplot(smoker[target], bins=20, label=target, rug=False,
kde=True, ax = axes[count//5][count%5])
        count = count+1
```



Considering the histogram for 22 samples does hard to see the details, choose box plot to vitualize the data

```
plt.figure(figsize=[25,10])
sns.boxplot(data=smoker,palette="Set3")
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1703845f8>
```



Question 1 Quiz: Why normalization is required for microarray data?

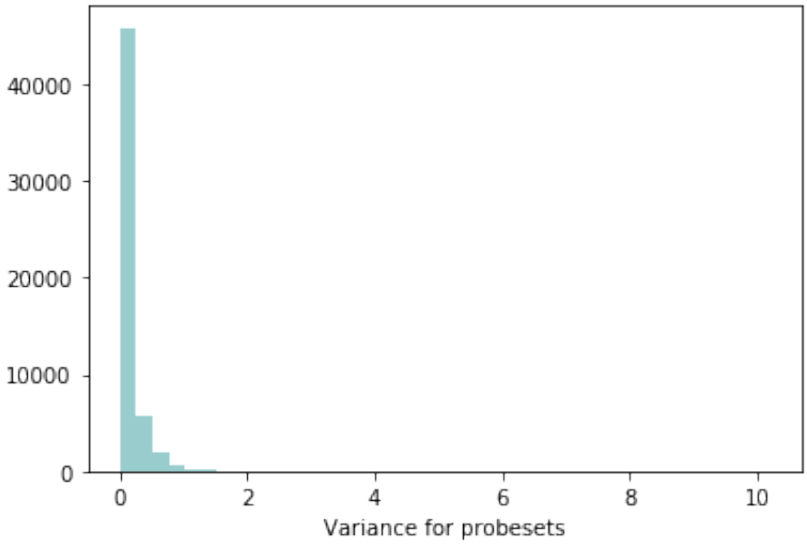
Anwser

1. Normalization is a process of eliminating the variations caused by differential labelling efficiency of teh two fluorescent dyes or different amounts of starting mRNA material in two samples. This process is suitable for this dataset: firstly, the dataset is consisted by different samples from different people sets, existing large possibility to have variation; secondly, as the reason we apply log2 normalization, it is obvious that the data varies large quantile in magnitude amount: some are thousands, some are units, which is abnormal, even after log transformation.
2. For such a large, random dataset, the result of the distribution should be closed to **Normal Distribution**.
   However, as we can observed from all the above figures that firstly, the central axis of the distribution **deviates from the central** from the range 0 to 16, most of the samples (global maximum value) accumulate at 3 to 5; secondly, the figure is **asystemitric**. This phenomenon is actually abnormal for a large dataset, which indicates there exists noise points or something else need to be refined by normalization.
3. The differences between two sets of samples(smokers and non-smokers) are hard to tell directly from the original figure, since the base number of the dataset is quite large, therefore even if the differences between samples seems small in some places, it could exist larger differential expression than we thought. In this perspective, normalization is also needed.

(e) Filter Genes

Before we perform unsupervised learning to explore the gene expression data, a common procedure is to filter out genes with low variance as they may simply represent noise. We can rank genes based on the variances across the dataset and select the 80% most variant probsets for use in the incoming analysis.

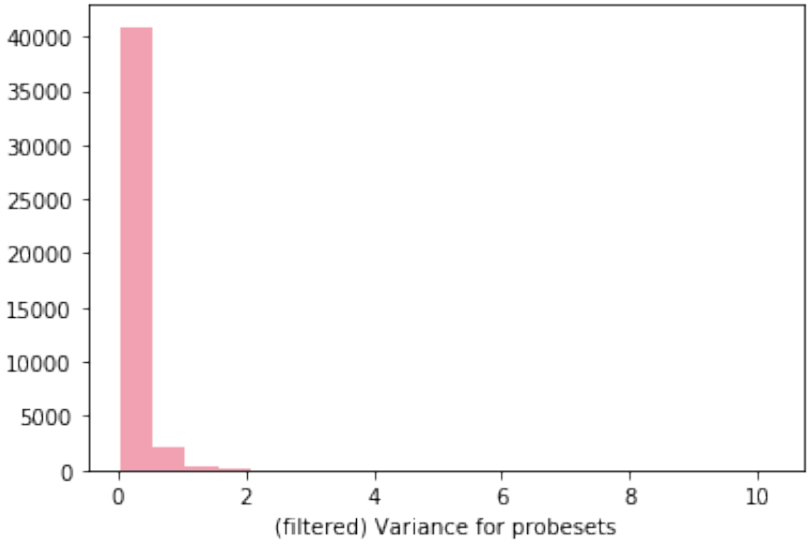
```
var_smoker = smoker.var(axis=1)
sns.distplot(var_smoker, bins=40, kde=False, color="#008080")
lab = plt.xlabel("Variance for probesets")
#plt.xlim([0, 25])
```




```
smoker_filt = smoker[pd.qcut(var_smoker, q=5, labels=False)  > 0]
print("Futher filtered dataset: %d filtered probsets, %d samples." %
smoker_filt.shape)
```

```
Futher filtered dataset: 43690 filtered probsets, 22 samples.
```

```
varsmoker_filt = smoker_filt.var(axis=1)
sns.distplot(varsmoker_filt, bins=20, kde=False, color="#dc143c")
lab = plt.xlabel("(filtered) Variance for probesets")
#plt.xlim(0, 25)
```



(f) Question 2 + 3: Wilcoxon Signed-Rank test, Paired Sample T-Test and Visualization

- Wilcoxon Signed-Rank Test
- Wilcoxon Signed-Rank Test Visualization
- Paired Sample T-Test
- Paired Sample T-Test Visualization

Divide two sets of samples according to labele.txt.
Notice that in order to make the balance to apply paired analysis, delete the final two samples in non_smoker_list.
The current result is 10 samples for each list.

```
non_smoker_list =
["GSM101095","GSM101096","GSM101097","GSM101098","GSM101099",

"GSM101100","GSM101101","GSM101102","GSM101103","GSM101104"]
smoker_list = ["GSM101107", "GSM101108", "GSM101109", "GSM101110",
"GSM101111",
               "GSM101112", "GSM101113", "GSM101114", "GSM101115",
"GSM101116"]
```

(a) Wilcoxon sign-rank test

- Setting P-Value is actually essential and here, we set it to be 0.006 instead of 0.05.
- Actually p-value = 0.05 is suitable for the analysis according to the original article: ***Pauwels RA, Buist AS, Calverley PM, Jenkins CR, Hurd SS (2001) Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease. NHLBI/WHO Global Initiative for Chronic Obstructive Lung Disease (GOLD) workshop summary. Am J Respir Crit Care Med 163:1256–1276***.
- However, we choose this p-value to generate a more strict result for the further dendrogram and clustering, since if the left data is too large, the dendrogram is hard to visualize without using *fastcluster* library.
- The p-value = 0.006 finally gives us **404 differential expressions**, which is acceptable.

```
significant_inds_wilcox, significant_pvals_wilcox = [], []
for ii in smoker_filt.index:
    pval = stats.wilcoxon(smoker_filt[non_smoker_list].loc[ii],
smoker_filt[smoker_list].loc[ii], alternative="two-sided").pvalue
    if pval < 0.006:
        significant_inds_wilcox.append(ii)
        significant_pvals_wilcox.append(pval)
        print("%s is significant probeset, with pvalue=%4f" % (ii, pval))
print(len(significant_inds_wilcox))
```

```
1552307_a_at is significant probeset, with pvalue=0.005062
1552497_a_at is significant probeset, with pvalue=0.005062
```

```
/Library/Frameworks/Python.framework/Versions/3.7/lib/python3.7/site-packages/scipy/stats/morestats.py:2863: UserWarning: Sample size too small for normal approximation.
  warnings.warn("Sample size too small for normal approximation.")
```

```
1552834_at is significant probeset, with pvalue=0.005062
1553172_at is significant probeset, with pvalue=0.005062
1553602_at is significant probeset, with pvalue=0.005062
1553709_a_at is significant probeset, with pvalue=0.005062
1553994_at is significant probeset, with pvalue=0.005062
1553995_a_at is significant probeset, with pvalue=0.005062
1555824_a_at is significant probeset, with pvalue=0.005062
1556467_at is significant probeset, with pvalue=0.005062
1556616_a_at is significant probeset, with pvalue=0.005062
1557038_s_at is significant probeset, with pvalue=0.005062
1557117_at is significant probeset, with pvalue=0.005062
1557158_s_at is significant probeset, with pvalue=0.005062
1557632_at is significant probeset, with pvalue=0.005062
1557719_at is significant probeset, with pvalue=0.005062
1557797_a_at is significant probeset, with pvalue=0.005062
1557965_at is significant probeset, with pvalue=0.005062
1558019_at is significant probeset, with pvalue=0.005062
1558703_at is significant probeset, with pvalue=0.005062
1558738_at is significant probeset, with pvalue=0.005062
1558868_a_at is significant probeset, with pvalue=0.005062
1559280_a_at is significant probeset, with pvalue=0.005062
1559814_at is significant probeset, with pvalue=0.005062
1560599_a_at is significant probeset, with pvalue=0.005062
1561002_at is significant probeset, with pvalue=0.005062
1562914_a_at is significant probeset, with pvalue=0.005062
1563189_at is significant probeset, with pvalue=0.005062
1563478_at is significant probeset, with pvalue=0.005062
1566123_at is significant probeset, with pvalue=0.005062
1566163_at is significant probeset, with pvalue=0.005062
1568699_at is significant probeset, with pvalue=0.005062
1568780_at is significant probeset, with pvalue=0.005062
1569917_at is significant probeset, with pvalue=0.005062
1570299_at is significant probeset, with pvalue=0.005062
200644_at is significant probeset, with pvalue=0.005062
200810_s_at is significant probeset, with pvalue=0.005062
200878_at is significant probeset, with pvalue=0.005062
201250_s_at is significant probeset, with pvalue=0.005062
201266_at is significant probeset, with pvalue=0.005062
201377_at is significant probeset, with pvalue=0.005062
201387_s_at is significant probeset, with pvalue=0.005062
201463_s_at is significant probeset, with pvalue=0.005062
201467_s_at is significant probeset, with pvalue=0.005062
201468_s_at is significant probeset, with pvalue=0.005062
```

```
201572_x_at is significant probeset, with pvalue=0.005062
201669_s_at is significant probeset, with pvalue=0.005062
201681_s_at is significant probeset, with pvalue=0.005062
201718_s_at is significant probeset, with pvalue=0.005062
201802_at is significant probeset, with pvalue=0.005062
201884_at is significant probeset, with pvalue=0.005062
201939_at is significant probeset, with pvalue=0.005062
201976_s_at is significant probeset, with pvalue=0.005062
202193_at is significant probeset, with pvalue=0.005062
202360_at is significant probeset, with pvalue=0.005062
202425_x_at is significant probeset, with pvalue=0.005062
202435_s_at is significant probeset, with pvalue=0.005062
202436_s_at is significant probeset, with pvalue=0.005062
202437_s_at is significant probeset, with pvalue=0.005062
202472_at is significant probeset, with pvalue=0.005062
202831_at is significant probeset, with pvalue=0.005062
202925_s_at is significant probeset, with pvalue=0.005062
203037_s_at is significant probeset, with pvalue=0.005062
203060_s_at is significant probeset, with pvalue=0.005062
203233_at is significant probeset, with pvalue=0.005062
203249_at is significant probeset, with pvalue=0.005062
203600_s_at is significant probeset, with pvalue=0.005062
203687_at is significant probeset, with pvalue=0.005062
203691_at is significant probeset, with pvalue=0.005062
203703_s_at is significant probeset, with pvalue=0.005062
203707_at is significant probeset, with pvalue=0.005062
203757_s_at is significant probeset, with pvalue=0.005062
203787_at is significant probeset, with pvalue=0.005062
203894_at is significant probeset, with pvalue=0.005062
203939_at is significant probeset, with pvalue=0.005062
204041_at is significant probeset, with pvalue=0.005062
204058_at is significant probeset, with pvalue=0.005062
204059_s_at is significant probeset, with pvalue=0.005062
204066_s_at is significant probeset, with pvalue=0.005062
204083_s_at is significant probeset, with pvalue=0.005062
204098_at is significant probeset, with pvalue=0.005062
204151_x_at is significant probeset, with pvalue=0.005062
204179_at is significant probeset, with pvalue=0.005062
204287_at is significant probeset, with pvalue=0.005062
204372_s_at is significant probeset, with pvalue=0.005062
204379_s_at is significant probeset, with pvalue=0.005062
204434_at is significant probeset, with pvalue=0.005062
204497_at is significant probeset, with pvalue=0.005062
204532_x_at is significant probeset, with pvalue=0.005062
204546_at is significant probeset, with pvalue=0.005062
204547_at is significant probeset, with pvalue=0.005062
204653_at is significant probeset, with pvalue=0.005062
204967_at is significant probeset, with pvalue=0.005062
205076_s_at is significant probeset, with pvalue=0.005062
```

```
205324_s_at is significant probeset, with pvalue=0.005062
205328_at is significant probeset, with pvalue=0.005062
205383_s_at is significant probeset, with pvalue=0.005062
205429_s_at is significant probeset, with pvalue=0.005062
205499_at is significant probeset, with pvalue=0.005062
205535_s_at is significant probeset, with pvalue=0.005062
205609_at is significant probeset, with pvalue=0.005062
205621_at is significant probeset, with pvalue=0.005062
205623_at is significant probeset, with pvalue=0.005062
205632_s_at is significant probeset, with pvalue=0.005062
205821_at is significant probeset, with pvalue=0.005062
206094_x_at is significant probeset, with pvalue=0.005062
206153_at is significant probeset, with pvalue=0.005062
206460_at is significant probeset, with pvalue=0.005062
206561_s_at is significant probeset, with pvalue=0.005062
206818_s_at is significant probeset, with pvalue=0.005062
206932_at is significant probeset, with pvalue=0.005062
207096_at is significant probeset, with pvalue=0.005062
207126_x_at is significant probeset, with pvalue=0.005062
207180_s_at is significant probeset, with pvalue=0.005062
207367_at is significant probeset, with pvalue=0.005062
207414_s_at is significant probeset, with pvalue=0.005062
207469_s_at is significant probeset, with pvalue=0.005062
207541_s_at is significant probeset, with pvalue=0.005062
207574_s_at is significant probeset, with pvalue=0.005062
207830_s_at is significant probeset, with pvalue=0.005062
208091_s_at is significant probeset, with pvalue=0.005062
208596_s_at is significant probeset, with pvalue=0.005062
208680_at is significant probeset, with pvalue=0.005062
208700_s_at is significant probeset, with pvalue=0.005062
209043_at is significant probeset, with pvalue=0.005062
209160_at is significant probeset, with pvalue=0.005062
209205_s_at is significant probeset, with pvalue=0.005062
209213_at is significant probeset, with pvalue=0.005062
209270_at is significant probeset, with pvalue=0.005062
209382_at is significant probeset, with pvalue=0.005062
209448_at is significant probeset, with pvalue=0.005062
209460_at is significant probeset, with pvalue=0.005062
209615_s_at is significant probeset, with pvalue=0.005062
209699_x_at is significant probeset, with pvalue=0.005062
209737_at is significant probeset, with pvalue=0.005062
210160_at is significant probeset, with pvalue=0.005062
210166_at is significant probeset, with pvalue=0.005062
210239_at is significant probeset, with pvalue=0.005062
210445_at is significant probeset, with pvalue=0.005062
210505_at is significant probeset, with pvalue=0.005062
210519_s_at is significant probeset, with pvalue=0.005062
210558_at is significant probeset, with pvalue=0.005062
210963_s_at is significant probeset, with pvalue=0.005062
```

```
211006_s_at is significant probeset, with pvalue=0.005062
211018_at is significant probeset, with pvalue=0.005062
211056_s_at is significant probeset, with pvalue=0.005062
211628_x_at is significant probeset, with pvalue=0.005062
211653_x_at is significant probeset, with pvalue=0.005062
211657_at is significant probeset, with pvalue=0.005062
211774_s_at is significant probeset, with pvalue=0.005062
211778_s_at is significant probeset, with pvalue=0.005062
212281_s_at is significant probeset, with pvalue=0.005062
212282_at is significant probeset, with pvalue=0.005062
212323_s_at is significant probeset, with pvalue=0.005062
212326_at is significant probeset, with pvalue=0.005062
212399_s_at is significant probeset, with pvalue=0.005062
212419_at is significant probeset, with pvalue=0.005062
212429_s_at is significant probeset, with pvalue=0.005062
212496_s_at is significant probeset, with pvalue=0.005062
212590_at is significant probeset, with pvalue=0.005062
212838_at is significant probeset, with pvalue=0.005062
212914_at is significant probeset, with pvalue=0.005062
212916_at is significant probeset, with pvalue=0.005062
213182_x_at is significant probeset, with pvalue=0.005062
213223_at is significant probeset, with pvalue=0.005062
213240_s_at is significant probeset, with pvalue=0.005062
213302_at is significant probeset, with pvalue=0.005062
213348_at is significant probeset, with pvalue=0.005062
213390_at is significant probeset, with pvalue=0.005062
213479_at is significant probeset, with pvalue=0.005062
213601_at is significant probeset, with pvalue=0.005062
213629_x_at is significant probeset, with pvalue=0.005062
213685_at is significant probeset, with pvalue=0.005062
213687_s_at is significant probeset, with pvalue=0.005062
213794_s_at is significant probeset, with pvalue=0.005062
213836_s_at is significant probeset, with pvalue=0.005062
214420_s_at is significant probeset, with pvalue=0.005062
214575_s_at is significant probeset, with pvalue=0.005062
214579_at is significant probeset, with pvalue=0.005062
214739_at is significant probeset, with pvalue=0.005062
214765_s_at is significant probeset, with pvalue=0.005062
214920_at is significant probeset, with pvalue=0.005062
215125_s_at is significant probeset, with pvalue=0.005062
215246_at is significant probeset, with pvalue=0.005062
215766_at is significant probeset, with pvalue=0.005062
215790_at is significant probeset, with pvalue=0.005062
216346_at is significant probeset, with pvalue=0.005062
216594_x_at is significant probeset, with pvalue=0.005062
216742_at is significant probeset, with pvalue=0.005062
216894_x_at is significant probeset, with pvalue=0.005062
217182_at is significant probeset, with pvalue=0.005062
217526_at is significant probeset, with pvalue=0.005062
```

```
217551_at is significant probeset, with pvalue=0.005062
217626_at is significant probeset, with pvalue=0.005062
217775_s_at is significant probeset, with pvalue=0.005062
218229_s_at is significant probeset, with pvalue=0.005062
218398_at is significant probeset, with pvalue=0.005062
218412_s_at is significant probeset, with pvalue=0.005062
218418_s_at is significant probeset, with pvalue=0.005062
218455_at is significant probeset, with pvalue=0.005062
218626_at is significant probeset, with pvalue=0.005062
218638_s_at is significant probeset, with pvalue=0.005062
218647_s_at is significant probeset, with pvalue=0.005062
218676_s_at is significant probeset, with pvalue=0.005062
218684_at is significant probeset, with pvalue=0.005062
218741_at is significant probeset, with pvalue=0.005062
218771_at is significant probeset, with pvalue=0.005062
218820_at is significant probeset, with pvalue=0.005062
218858_at is significant probeset, with pvalue=0.005062
218950_at is significant probeset, with pvalue=0.005062
219049_at is significant probeset, with pvalue=0.005062
219060_at is significant probeset, with pvalue=0.005062
219123_at is significant probeset, with pvalue=0.005062
219383_at is significant probeset, with pvalue=0.005062
219405_at is significant probeset, with pvalue=0.005062
219410_at is significant probeset, with pvalue=0.005062
219450_at is significant probeset, with pvalue=0.005062
219534_x_at is significant probeset, with pvalue=0.005062
219563_at is significant probeset, with pvalue=0.005062
219641_at is significant probeset, with pvalue=0.005062
219685_at is significant probeset, with pvalue=0.005062
219743_at is significant probeset, with pvalue=0.005062
219765_at is significant probeset, with pvalue=0.005062
219928_s_at is significant probeset, with pvalue=0.005062
219944_at is significant probeset, with pvalue=0.005062
219958_at is significant probeset, with pvalue=0.005062
219966_x_at is significant probeset, with pvalue=0.005062
220003_at is significant probeset, with pvalue=0.005062
220066_at is significant probeset, with pvalue=0.005062
220177_s_at is significant probeset, with pvalue=0.005062
220197_at is significant probeset, with pvalue=0.005062
220610_s_at is significant probeset, with pvalue=0.005062
221016_s_at is significant probeset, with pvalue=0.005062
221096_s_at is significant probeset, with pvalue=0.005062
221175_at is significant probeset, with pvalue=0.005062
221288_at is significant probeset, with pvalue=0.005062
221538_s_at is significant probeset, with pvalue=0.005062
221567_at is significant probeset, with pvalue=0.005062
221575_at is significant probeset, with pvalue=0.005062
221619_s_at is significant probeset, with pvalue=0.005062
221636_s_at is significant probeset, with pvalue=0.005062
```

```
221675_s_at is significant probeset, with pvalue=0.005062
221867_at is significant probeset, with pvalue=0.005062
221909_at is significant probeset, with pvalue=0.005062
222072_at is significant probeset, with pvalue=0.005062
222088_s_at is significant probeset, with pvalue=0.005062
222455_s_at is significant probeset, with pvalue=0.005062
222513_s_at is significant probeset, with pvalue=0.005062
222537_s_at is significant probeset, with pvalue=0.005062
222561_at is significant probeset, with pvalue=0.005062
222757_s_at is significant probeset, with pvalue=0.005062
222921_s_at is significant probeset, with pvalue=0.005062
223040_at is significant probeset, with pvalue=0.005062
223114_at is significant probeset, with pvalue=0.005062
223120_at is significant probeset, with pvalue=0.005062
223593_at is significant probeset, with pvalue=0.005062
223714_at is significant probeset, with pvalue=0.005062
223792_at is significant probeset, with pvalue=0.005062
223822_at is significant probeset, with pvalue=0.005062
224279_s_at is significant probeset, with pvalue=0.005062
224325_at is significant probeset, with pvalue=0.005062
224681_at is significant probeset, with pvalue=0.005062
225005_at is significant probeset, with pvalue=0.005062
225016_at is significant probeset, with pvalue=0.005062
225116_at is significant probeset, with pvalue=0.005062
225117_at is significant probeset, with pvalue=0.005062
225252_at is significant probeset, with pvalue=0.005062
225296_at is significant probeset, with pvalue=0.005062
225337_at is significant probeset, with pvalue=0.005062
225395_s_at is significant probeset, with pvalue=0.005062
225402_at is significant probeset, with pvalue=0.005062
225567_at is significant probeset, with pvalue=0.005062
225606_at is significant probeset, with pvalue=0.005062
225637_at is significant probeset, with pvalue=0.005062
225704_at is significant probeset, with pvalue=0.005062
225800_at is significant probeset, with pvalue=0.005062
225843_at is significant probeset, with pvalue=0.005062
225962_at is significant probeset, with pvalue=0.005062
226032_at is significant probeset, with pvalue=0.005062
226093_at is significant probeset, with pvalue=0.005062
226116_at is significant probeset, with pvalue=0.005062
226194_at is significant probeset, with pvalue=0.005062
226206_at is significant probeset, with pvalue=0.005062
226267_at is significant probeset, with pvalue=0.005062
226443_at is significant probeset, with pvalue=0.005062
226509_at is significant probeset, with pvalue=0.005062
226780_s_at is significant probeset, with pvalue=0.005062
227084_at is significant probeset, with pvalue=0.005062
227155_at is significant probeset, with pvalue=0.005062
227168_at is significant probeset, with pvalue=0.005062
```

```
227197_at is significant probeset, with pvalue=0.005062
227334_at is significant probeset, with pvalue=0.005062
227405_s_at is significant probeset, with pvalue=0.005062
227475_at is significant probeset, with pvalue=0.005062
227515_at is significant probeset, with pvalue=0.005062
227516_at is significant probeset, with pvalue=0.005062
227558_at is significant probeset, with pvalue=0.005062
227572_at is significant probeset, with pvalue=0.005062
227593_at is significant probeset, with pvalue=0.005062
227615_at is significant probeset, with pvalue=0.005062
227702_at is significant probeset, with pvalue=0.005062
227742_at is significant probeset, with pvalue=0.005062
228055_at is significant probeset, with pvalue=0.005062
228093_at is significant probeset, with pvalue=0.005062
228303_at is significant probeset, with pvalue=0.005062
228412_at is significant probeset, with pvalue=0.005062
228461_at is significant probeset, with pvalue=0.005062
228490_at is significant probeset, with pvalue=0.005062
228710_at is significant probeset, with pvalue=0.005062
228738_at is significant probeset, with pvalue=0.005062
228811_at is significant probeset, with pvalue=0.005062
228855_at is significant probeset, with pvalue=0.005062
228856_at is significant probeset, with pvalue=0.005062
228899_at is significant probeset, with pvalue=0.005062
229157_at is significant probeset, with pvalue=0.005062
229158_at is significant probeset, with pvalue=0.005062
229175_at is significant probeset, with pvalue=0.005062
229265_at is significant probeset, with pvalue=0.005062
229309_at is significant probeset, with pvalue=0.005062
229407_at is significant probeset, with pvalue=0.005062
229537_at is significant probeset, with pvalue=0.005062
229566_at is significant probeset, with pvalue=0.005062
229852_at is significant probeset, with pvalue=0.005062
229964_at is significant probeset, with pvalue=0.005062
230054_at is significant probeset, with pvalue=0.005062
230100_x_at is significant probeset, with pvalue=0.005062
230311_s_at is significant probeset, with pvalue=0.005062
230433_at is significant probeset, with pvalue=0.005062
230488_s_at is significant probeset, with pvalue=0.005062
230489_at is significant probeset, with pvalue=0.005062
230747_s_at is significant probeset, with pvalue=0.005062
230776_at is significant probeset, with pvalue=0.005062
230782_at is significant probeset, with pvalue=0.005062
230849_at is significant probeset, with pvalue=0.005062
230857_s_at is significant probeset, with pvalue=0.005062
230888_at is significant probeset, with pvalue=0.005062
231487_at is significant probeset, with pvalue=0.005062
231800_s_at is significant probeset, with pvalue=0.005062
231835_at is significant probeset, with pvalue=0.005062
```

```
231842_at is significant probeset, with pvalue=0.005062
231845_at is significant probeset, with pvalue=0.005062
231901_at is significant probeset, with pvalue=0.005062
231928_at is significant probeset, with pvalue=0.005062
232059_at is significant probeset, with pvalue=0.005062
232478_at is significant probeset, with pvalue=0.005062
232654_s_at is significant probeset, with pvalue=0.005062
232766_at is significant probeset, with pvalue=0.005062
232891_at is significant probeset, with pvalue=0.005062
233852_at is significant probeset, with pvalue=0.005062
234073_at is significant probeset, with pvalue=0.005062
234211_at is significant probeset, with pvalue=0.005062
234312_s_at is significant probeset, with pvalue=0.005062
234317_s_at is significant probeset, with pvalue=0.005062
234329_at is significant probeset, with pvalue=0.005062
234561_at is significant probeset, with pvalue=0.005062
235174_s_at is significant probeset, with pvalue=0.005062
235350_at is significant probeset, with pvalue=0.005062
235647_at is significant probeset, with pvalue=0.005062
235727_at is significant probeset, with pvalue=0.005062
235793_at is significant probeset, with pvalue=0.005062
235804_at is significant probeset, with pvalue=0.005062
235845_at is significant probeset, with pvalue=0.005062
235920_at is significant probeset, with pvalue=0.005062
236124_at is significant probeset, with pvalue=0.005062
236132_at is significant probeset, with pvalue=0.005062
236465_at is significant probeset, with pvalue=0.005062
236656_s_at is significant probeset, with pvalue=0.005062
237351_at is significant probeset, with pvalue=0.005062
238755_at is significant probeset, with pvalue=0.005062
238986_at is significant probeset, with pvalue=0.005062
238999_at is significant probeset, with pvalue=0.005062
239021_at is significant probeset, with pvalue=0.005062
239093_at is significant probeset, with pvalue=0.005062
239229_at is significant probeset, with pvalue=0.005062
239283_at is significant probeset, with pvalue=0.005062
240382_at is significant probeset, with pvalue=0.005062
240454_at is significant probeset, with pvalue=0.005062
240699_at is significant probeset, with pvalue=0.005062
240785_at is significant probeset, with pvalue=0.005062
240788_at is significant probeset, with pvalue=0.005062
240869_at is significant probeset, with pvalue=0.005062
240899_at is significant probeset, with pvalue=0.005062
241233_x_at is significant probeset, with pvalue=0.005062
241315_at is significant probeset, with pvalue=0.005062
241418_at is significant probeset, with pvalue=0.005062
241764_at is significant probeset, with pvalue=0.005062
241877_at is significant probeset, with pvalue=0.005062
241890_at is significant probeset, with pvalue=0.005062
```

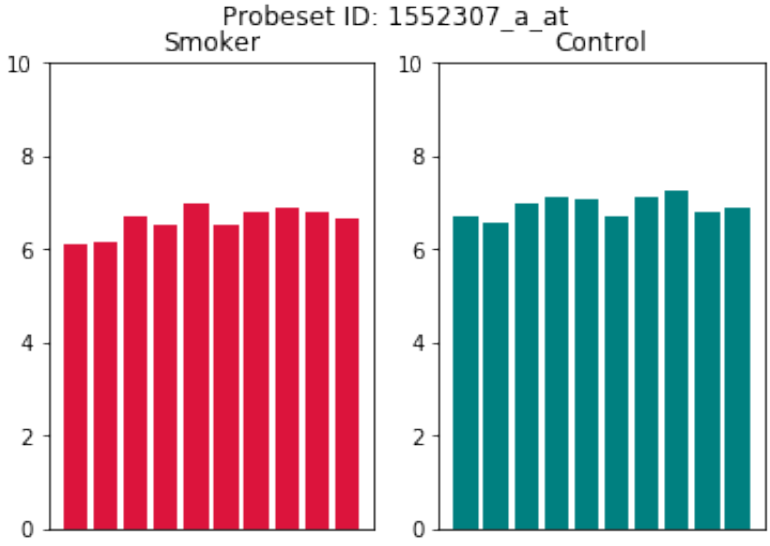
```
241950_at is significant probeset, with pvalue=0.005062
242452_at is significant probeset, with pvalue=0.005062
242478_at is significant probeset, with pvalue=0.005062
243594_x_at is significant probeset, with pvalue=0.005062
244519_at is significant probeset, with pvalue=0.005062
244600_at is significant probeset, with pvalue=0.005062
244654_at is significant probeset, with pvalue=0.005062
244677_at is significant probeset, with pvalue=0.005062
37586_at is significant probeset, with pvalue=0.005062
40284_at is significant probeset, with pvalue=0.005062
41577_at is significant probeset, with pvalue=0.005062
42361_g_at is significant probeset, with pvalue=0.005062
45526_g_at is significant probeset, with pvalue=0.005062
55081_at is significant probeset, with pvalue=0.005062
823_at is significant probeset, with pvalue=0.005062
91617_at is significant probeset, with pvalue=0.005062
404
```

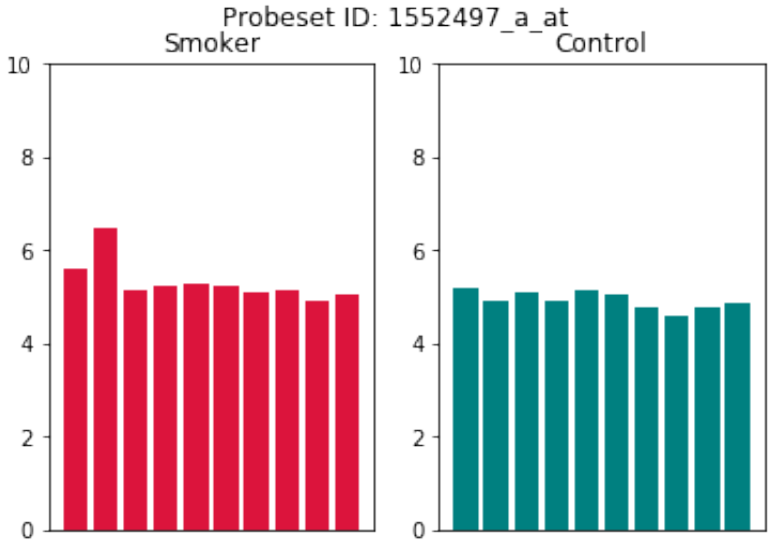
(b) Wilcoxon sign-rank test Vistualization

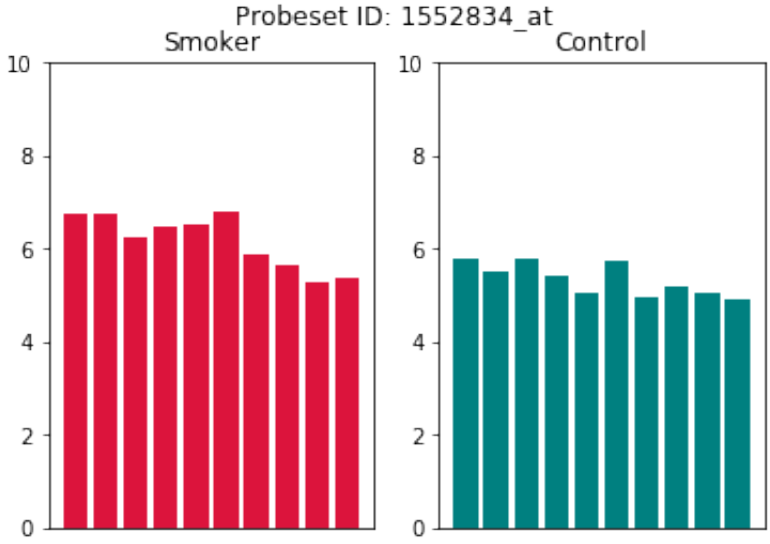
select Gene: '1552307_a_at','1552497_a_at','1552834_at'

- Bar Plot
- Scatter Plot

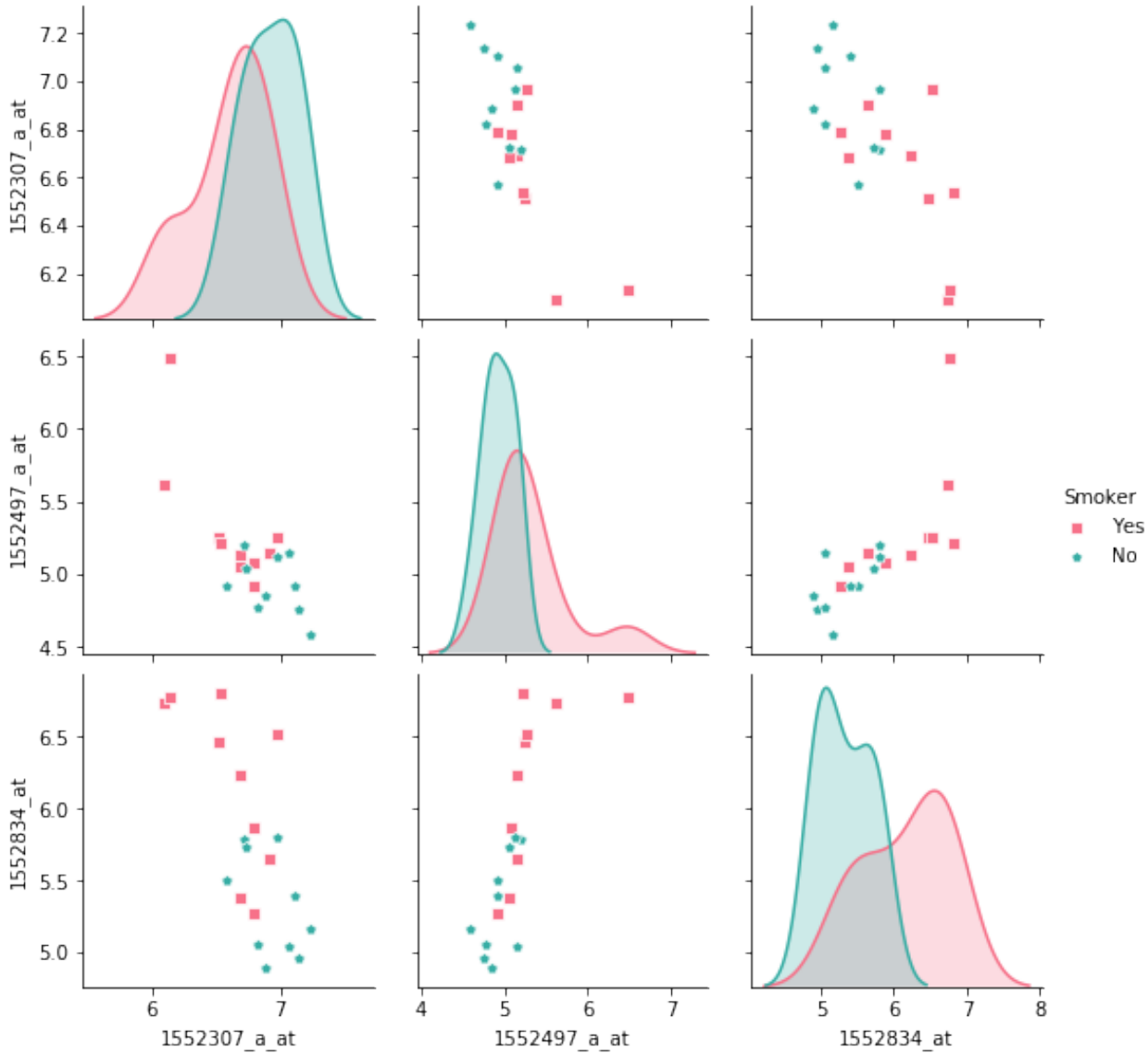
```
select_w_list= ['1552307_a_at','1552497_a_at','1552834_at']
smoker_wilcox = smoker_filt[non_smoker_list + smoker_list].loc[select_w_list]
for ii in smoker_wilcox.index:
    fig = plt.figure()
    ax1 = fig.add_subplot(121)
    ax1.bar(smoker_list, smoker_wilcox[smoker_list].loc[ii], color="#dc143c")
    plt.xticks(''); plt.title("Smoker"); plt.ylim(0, 10)
    ax2 = fig.add_subplot(122)
    ax2.bar(smoker_list, smoker_wilcox[non_smoker_list].loc[ii],
color="#008080")
    plt.xticks(''); plt.title("Control"); plt.ylim(0, 10)
    fig.suptitle("Probeset ID: %s" % ii)
```


Probeset ID: 1552307_a_at
Smoker
Control
10
8
6
4
2
0


Probeset ID: 1552497_a_at
Smoker
Control
10
8
6
4
2
0


Probeset ID: 1552834_at
Smoker
Control
10
8
6
4
2
0

```
# Scatter plots combining the probesets
df_scatter_s = smoker_wilcox[smoker_list].T
df_scatter_s['Smoker'] = 'Yes'
df_scatter_c = smoker_wilcox[non_smoker_list].T
df_scatter_c['Smoker'] = 'No'
df_scatter = pd.concat([df_scatter_s, df_scatter_c], axis=0)
# print(df_scatter)
# print(df_scatter.index)
# Pairplot
pp = sns.pairplot(df_scatter, hue='Smoker', diag_kind='auto', markers=['s',
'p'], palette="husl")
```



(c) T-test

- Setting P-Value is actually essential and here, we set it to be 0.003 instead of 0.05.
- Actually p-value = 0.05 is suitable for the analysis according to the original article: ***Pauwels RA, Buist AS, Calverley PM, Jenkins CR, Hurd SS (2001) Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease. NHLBI/WHO Global Initiative for Chronic Obstructive Lung Disease (GOLD) workshop summary. Am J Respir Crit Care Med 163:1256–1276***.

- However, we choose this p-value to generate a more strict result for the further dendrogram and clustering, since if the left data is too large, the dendrogram is hard to visualize without using *fastcluster* library.
- The p-value = 0.006 finally gives us **642 differential expressions**, which is acceptable but it would be better to use *fastcluster*.
- Notice the p-value compared to the Wilcoxon sign-rank test is even smaller with more differential expressions, which reflects Wilcoxon sign-rank test might be a more strict method in this problem

```
significant_inds_ttest, significant_pvals_ttest = [], []
for ii in smoker_filt.index:
    pval = stats.ttest_rel(smoker_filt[non_smoker_list].loc[ii],
smoker_filt[smoker_list].loc[ii]).pvalue
    if pval < 0.003:
        significant_inds_ttest.append(ii)
        significant_pvals_ttest.append(pval)
        print("%s is significant probeset, with pvalue=%4f" % (ii, pval))
print(len(significant_inds_ttest))
```

```
1552307_a_at is significant probeset, with pvalue=0.000734
1552833_at is significant probeset, with pvalue=0.001770
1552834_at is significant probeset, with pvalue=0.000117
1553172_at is significant probeset, with pvalue=0.000126
1553602_at is significant probeset, with pvalue=0.000005
1553704_x_at is significant probeset, with pvalue=0.000575
1553709_a_at is significant probeset, with pvalue=0.000588
1553729_s_at is significant probeset, with pvalue=0.000685
1553961_s_at is significant probeset, with pvalue=0.002265
1553994_at is significant probeset, with pvalue=0.000737
1553995_a_at is significant probeset, with pvalue=0.000100
1554085_at is significant probeset, with pvalue=0.002918
1554168_a_at is significant probeset, with pvalue=0.001449
1554182_at is significant probeset, with pvalue=0.002013
1554190_s_at is significant probeset, with pvalue=0.000953
1555095_at is significant probeset, with pvalue=0.002860
1555824_a_at is significant probeset, with pvalue=0.000334
1555854_at is significant probeset, with pvalue=0.001114
1555886_at is significant probeset, with pvalue=0.000580
1556082_a_at is significant probeset, with pvalue=0.001537
1557038_s_at is significant probeset, with pvalue=0.000272
1557117_at is significant probeset, with pvalue=0.001640
1557136_at is significant probeset, with pvalue=0.002292
1557158_s_at is significant probeset, with pvalue=0.000133
1557585_at is significant probeset, with pvalue=0.001565
1557632_at is significant probeset, with pvalue=0.001766
1557658_at is significant probeset, with pvalue=0.001912
1557681_s_at is significant probeset, with pvalue=0.002270
1557965_at is significant probeset, with pvalue=0.000615
```

```
1558019_at is significant probeset, with pvalue=0.001697
1558586_at is significant probeset, with pvalue=0.001126
1558703_at is significant probeset, with pvalue=0.000055
1558738_at is significant probeset, with pvalue=0.001819
1558868_a_at is significant probeset, with pvalue=0.000213
1559072_a_at is significant probeset, with pvalue=0.000731
1559280_a_at is significant probeset, with pvalue=0.001556
1559922_at is significant probeset, with pvalue=0.002069
1559946_s_at is significant probeset, with pvalue=0.002792
1560599_a_at is significant probeset, with pvalue=0.001767
1561002_at is significant probeset, with pvalue=0.002369
1562914_a_at is significant probeset, with pvalue=0.002568
1563189_at is significant probeset, with pvalue=0.002810
1566123_at is significant probeset, with pvalue=0.000024
1568699_at is significant probeset, with pvalue=0.001946
1568780_at is significant probeset, with pvalue=0.001240
1570299_at is significant probeset, with pvalue=0.002415
200019_s_at is significant probeset, with pvalue=0.002085
200036_s_at is significant probeset, with pvalue=0.002951
200062_s_at is significant probeset, with pvalue=0.002878
200088_x_at is significant probeset, with pvalue=0.001943
200644_at is significant probeset, with pvalue=0.001731
200748_s_at is significant probeset, with pvalue=0.000886
200783_s_at is significant probeset, with pvalue=0.002121
200804_at is significant probeset, with pvalue=0.001769
200810_s_at is significant probeset, with pvalue=0.000024
200872_at is significant probeset, with pvalue=0.000925
200875_s_at is significant probeset, with pvalue=0.001774
200878_at is significant probeset, with pvalue=0.000417
201054_at is significant probeset, with pvalue=0.001824
201250_s_at is significant probeset, with pvalue=0.000151
201266_at is significant probeset, with pvalue=0.000547
201272_at is significant probeset, with pvalue=0.002706
201377_at is significant probeset, with pvalue=0.000549
201387_s_at is significant probeset, with pvalue=0.000001
201463_s_at is significant probeset, with pvalue=0.000017
201467_s_at is significant probeset, with pvalue=0.000007
201468_s_at is significant probeset, with pvalue=0.000011
201487_at is significant probeset, with pvalue=0.000822
201566_x_at is significant probeset, with pvalue=0.001529
201572_x_at is significant probeset, with pvalue=0.000361
201591_s_at is significant probeset, with pvalue=0.002431
201669_s_at is significant probeset, with pvalue=0.002586
201681_s_at is significant probeset, with pvalue=0.001575
201707_at is significant probeset, with pvalue=0.000469
201718_s_at is significant probeset, with pvalue=0.000046
201719_s_at is significant probeset, with pvalue=0.000328
201802_at is significant probeset, with pvalue=0.002553
201884_at is significant probeset, with pvalue=0.001826
```

```
201934_at is significant probeset, with pvalue=0.002607
201939_at is significant probeset, with pvalue=0.002726
201976_s_at is significant probeset, with pvalue=0.001143
202084_s_at is significant probeset, with pvalue=0.000582
202153_s_at is significant probeset, with pvalue=0.002711
202230_s_at is significant probeset, with pvalue=0.002425
202360_at is significant probeset, with pvalue=0.000896
202425_x_at is significant probeset, with pvalue=0.002243
202435_s_at is significant probeset, with pvalue=0.000368
202436_s_at is significant probeset, with pvalue=0.000153
202437_s_at is significant probeset, with pvalue=0.000301
202472_at is significant probeset, with pvalue=0.000007
202495_at is significant probeset, with pvalue=0.000707
202560_s_at is significant probeset, with pvalue=0.001307
202577_s_at is significant probeset, with pvalue=0.001946
202729_s_at is significant probeset, with pvalue=0.001639
202831_at is significant probeset, with pvalue=0.000023
202860_at is significant probeset, with pvalue=0.001078
202888_s_at is significant probeset, with pvalue=0.001294
202923_s_at is significant probeset, with pvalue=0.002923
202925_s_at is significant probeset, with pvalue=0.001529
203037_s_at is significant probeset, with pvalue=0.000177
203060_s_at is significant probeset, with pvalue=0.001646
203062_s_at is significant probeset, with pvalue=0.002679
203192_at is significant probeset, with pvalue=0.000388
203233_at is significant probeset, with pvalue=0.000339
203286_at is significant probeset, with pvalue=0.001638
203353_s_at is significant probeset, with pvalue=0.001498
203390_s_at is significant probeset, with pvalue=0.000461
203468_at is significant probeset, with pvalue=0.000645
203498_at is significant probeset, with pvalue=0.001396
203600_s_at is significant probeset, with pvalue=0.000213
203687_at is significant probeset, with pvalue=0.000951
203703_s_at is significant probeset, with pvalue=0.000065
203707_at is significant probeset, with pvalue=0.000521
203757_s_at is significant probeset, with pvalue=0.002123
203787_at is significant probeset, with pvalue=0.000989
203818_s_at is significant probeset, with pvalue=0.000848
203851_at is significant probeset, with pvalue=0.001158
203865_s_at is significant probeset, with pvalue=0.002873
203894_at is significant probeset, with pvalue=0.001330
203924_at is significant probeset, with pvalue=0.000914
203925_at is significant probeset, with pvalue=0.001584
203939_at is significant probeset, with pvalue=0.000061
203956_at is significant probeset, with pvalue=0.000288
204000_at is significant probeset, with pvalue=0.002933
204041_at is significant probeset, with pvalue=0.000969
204058_at is significant probeset, with pvalue=0.000738
204059_s_at is significant probeset, with pvalue=0.000013
```

```
204066_s_at is significant probeset, with pvalue=0.000066
204083_s_at is significant probeset, with pvalue=0.001742
204098_at is significant probeset, with pvalue=0.000283
204151_x_at is significant probeset, with pvalue=0.000035
204173_at is significant probeset, with pvalue=0.001110
204179_at is significant probeset, with pvalue=0.000436
204287_at is significant probeset, with pvalue=0.000048
204341_at is significant probeset, with pvalue=0.001365
204366_s_at is significant probeset, with pvalue=0.001076
204367_at is significant probeset, with pvalue=0.000889
204372_s_at is significant probeset, with pvalue=0.001542
204379_s_at is significant probeset, with pvalue=0.000800
204400_at is significant probeset, with pvalue=0.001819
204434_at is significant probeset, with pvalue=0.000466
204529_s_at is significant probeset, with pvalue=0.002972
204532_x_at is significant probeset, with pvalue=0.000001
204546_at is significant probeset, with pvalue=0.000577
204547_at is significant probeset, with pvalue=0.001859
204607_at is significant probeset, with pvalue=0.000802
204675_at is significant probeset, with pvalue=0.001934
204967_at is significant probeset, with pvalue=0.002955
204970_s_at is significant probeset, with pvalue=0.001086
205221_at is significant probeset, with pvalue=0.001127
205241_at is significant probeset, with pvalue=0.002041
205267_at is significant probeset, with pvalue=0.001203
205324_s_at is significant probeset, with pvalue=0.000728
205328_at is significant probeset, with pvalue=0.000023
205379_at is significant probeset, with pvalue=0.000796
205383_s_at is significant probeset, with pvalue=0.000393
205429_s_at is significant probeset, with pvalue=0.001555
205499_at is significant probeset, with pvalue=0.000060
205513_at is significant probeset, with pvalue=0.002036
205535_s_at is significant probeset, with pvalue=0.000361
205566_at is significant probeset, with pvalue=0.000681
205621_at is significant probeset, with pvalue=0.000107
205623_at is significant probeset, with pvalue=0.000028
205632_s_at is significant probeset, with pvalue=0.000157
205741_s_at is significant probeset, with pvalue=0.001614
205821_at is significant probeset, with pvalue=0.000646
205933_at is significant probeset, with pvalue=0.002758
206000_at is significant probeset, with pvalue=0.000657
206082_at is significant probeset, with pvalue=0.002402
206094_x_at is significant probeset, with pvalue=0.000004
206153_at is significant probeset, with pvalue=0.001095
206170_at is significant probeset, with pvalue=0.002917
206263_at is significant probeset, with pvalue=0.001723
206460_at is significant probeset, with pvalue=0.000121
206472_s_at is significant probeset, with pvalue=0.001756
206492_at is significant probeset, with pvalue=0.002420
```

```
206561_s_at is significant probeset, with pvalue=0.000006
206670_s_at is significant probeset, with pvalue=0.001130
207096_at is significant probeset, with pvalue=0.000368
207126_x_at is significant probeset, with pvalue=0.000002
207157_s_at is significant probeset, with pvalue=0.002445
207180_s_at is significant probeset, with pvalue=0.000000
207367_at is significant probeset, with pvalue=0.000400
207469_s_at is significant probeset, with pvalue=0.000049
207517_at is significant probeset, with pvalue=0.001979
207541_s_at is significant probeset, with pvalue=0.000308
207574_s_at is significant probeset, with pvalue=0.000853
207618_s_at is significant probeset, with pvalue=0.000648
207753_at is significant probeset, with pvalue=0.002744
207830_s_at is significant probeset, with pvalue=0.000601
208136_s_at is significant probeset, with pvalue=0.002838
208424_s_at is significant probeset, with pvalue=0.002278
208517_x_at is significant probeset, with pvalue=0.000958
208596_s_at is significant probeset, with pvalue=0.000004
208636_at is significant probeset, with pvalue=0.002922
208680_at is significant probeset, with pvalue=0.000791
208700_s_at is significant probeset, with pvalue=0.000307
208984_x_at is significant probeset, with pvalue=0.001493
209043_at is significant probeset, with pvalue=0.000824
209153_s_at is significant probeset, with pvalue=0.002674
209160_at is significant probeset, with pvalue=0.000044
209161_at is significant probeset, with pvalue=0.000862
209213_at is significant probeset, with pvalue=0.000064
209343_at is significant probeset, with pvalue=0.001308
209355_s_at is significant probeset, with pvalue=0.002716
209382_at is significant probeset, with pvalue=0.000219
209392_at is significant probeset, with pvalue=0.002559
209448_at is significant probeset, with pvalue=0.000104
209615_s_at is significant probeset, with pvalue=0.000589
209625_at is significant probeset, with pvalue=0.001439
209667_at is significant probeset, with pvalue=0.001664
209699_x_at is significant probeset, with pvalue=0.000004
209781_s_at is significant probeset, with pvalue=0.001707
209841_s_at is significant probeset, with pvalue=0.002697
209921_at is significant probeset, with pvalue=0.001230
210115_at is significant probeset, with pvalue=0.001740
210160_at is significant probeset, with pvalue=0.002286
210166_at is significant probeset, with pvalue=0.000306
210445_at is significant probeset, with pvalue=0.000251
210505_at is significant probeset, with pvalue=0.000012
210519_s_at is significant probeset, with pvalue=0.000003
210524_x_at is significant probeset, with pvalue=0.000954
210617_at is significant probeset, with pvalue=0.001660
210760_x_at is significant probeset, with pvalue=0.000301
210769_at is significant probeset, with pvalue=0.001307
```

```
210963_s_at is significant probeset, with pvalue=0.002514
211006_s_at is significant probeset, with pvalue=0.000427
211056_s_at is significant probeset, with pvalue=0.000886
211628_x_at is significant probeset, with pvalue=0.002917
211653_x_at is significant probeset, with pvalue=0.000020
211657_at is significant probeset, with pvalue=0.001215
211717_at is significant probeset, with pvalue=0.001740
211774_s_at is significant probeset, with pvalue=0.000343
211778_s_at is significant probeset, with pvalue=0.000466
211998_at is significant probeset, with pvalue=0.000869
212180_at is significant probeset, with pvalue=0.001619
212230_at is significant probeset, with pvalue=0.001070
212279_at is significant probeset, with pvalue=0.002178
212281_s_at is significant probeset, with pvalue=0.000142
212282_at is significant probeset, with pvalue=0.000863
212323_s_at is significant probeset, with pvalue=0.000341
212326_at is significant probeset, with pvalue=0.002521
212399_s_at is significant probeset, with pvalue=0.001215
212419_at is significant probeset, with pvalue=0.001746
212429_s_at is significant probeset, with pvalue=0.000141
212472_at is significant probeset, with pvalue=0.002261
212496_s_at is significant probeset, with pvalue=0.000597
212590_at is significant probeset, with pvalue=0.000487
212617_at is significant probeset, with pvalue=0.001062
212686_at is significant probeset, with pvalue=0.002292
212750_at is significant probeset, with pvalue=0.000826
212838_at is significant probeset, with pvalue=0.000494
212841_s_at is significant probeset, with pvalue=0.002459
212904_at is significant probeset, with pvalue=0.002260
212914_at is significant probeset, with pvalue=0.002033
212916_at is significant probeset, with pvalue=0.000240
212954_at is significant probeset, with pvalue=0.000438
213069_at is significant probeset, with pvalue=0.001126
213124_at is significant probeset, with pvalue=0.000424
213198_at is significant probeset, with pvalue=0.001404
213217_at is significant probeset, with pvalue=0.000739
213223_at is significant probeset, with pvalue=0.000014
213240_s_at is significant probeset, with pvalue=0.000738
213302_at is significant probeset, with pvalue=0.000116
213316_at is significant probeset, with pvalue=0.002338
213340_s_at is significant probeset, with pvalue=0.002691
213348_at is significant probeset, with pvalue=0.000121
213390_at is significant probeset, with pvalue=0.000129
213488_at is significant probeset, with pvalue=0.000494
213526_s_at is significant probeset, with pvalue=0.002749
213601_at is significant probeset, with pvalue=0.001396
213629_x_at is significant probeset, with pvalue=0.000484
213687_s_at is significant probeset, with pvalue=0.001518
213720_s_at is significant probeset, with pvalue=0.002851
```

```
213794_s_at is significant probeset, with pvalue=0.000242
213836_s_at is significant probeset, with pvalue=0.000230
214086_s_at is significant probeset, with pvalue=0.002657
214211_at is significant probeset, with pvalue=0.001097
214308_s_at is significant probeset, with pvalue=0.002807
214710_s_at is significant probeset, with pvalue=0.002641
214711_at is significant probeset, with pvalue=0.001362
214739_at is significant probeset, with pvalue=0.000104
215012_at is significant probeset, with pvalue=0.002846
215125_s_at is significant probeset, with pvalue=0.000003
215246_at is significant probeset, with pvalue=0.000176
215773_x_at is significant probeset, with pvalue=0.002537
215790_at is significant probeset, with pvalue=0.000334
215850_s_at is significant probeset, with pvalue=0.001355
216346_at is significant probeset, with pvalue=0.000659
216506_x_at is significant probeset, with pvalue=0.001842
216594_x_at is significant probeset, with pvalue=0.000072
216742_at is significant probeset, with pvalue=0.002825
216894_x_at is significant probeset, with pvalue=0.001063
217182_at is significant probeset, with pvalue=0.001556
217526_at is significant probeset, with pvalue=0.001226
217546_at is significant probeset, with pvalue=0.001020
217626_at is significant probeset, with pvalue=0.000261
217678_at is significant probeset, with pvalue=0.000919
217775_s_at is significant probeset, with pvalue=0.000747
217948_at is significant probeset, with pvalue=0.001726
217978_s_at is significant probeset, with pvalue=0.002133
218023_s_at is significant probeset, with pvalue=0.000245
218218_at is significant probeset, with pvalue=0.002079
218229_s_at is significant probeset, with pvalue=0.000284
218378_s_at is significant probeset, with pvalue=0.001065
218398_at is significant probeset, with pvalue=0.001368
218412_s_at is significant probeset, with pvalue=0.000707
218418_s_at is significant probeset, with pvalue=0.000446
218443_s_at is significant probeset, with pvalue=0.001416
218449_at is significant probeset, with pvalue=0.001271
218455_at is significant probeset, with pvalue=0.000050
218626_at is significant probeset, with pvalue=0.000254
218641_at is significant probeset, with pvalue=0.001884
218647_s_at is significant probeset, with pvalue=0.000104
218648_at is significant probeset, with pvalue=0.001354
218671_s_at is significant probeset, with pvalue=0.002970
218676_s_at is significant probeset, with pvalue=0.000398
218684_at is significant probeset, with pvalue=0.000280
218722_s_at is significant probeset, with pvalue=0.001124
218741_at is significant probeset, with pvalue=0.000814
218771_at is significant probeset, with pvalue=0.000283
218820_at is significant probeset, with pvalue=0.001554
218858_at is significant probeset, with pvalue=0.000393
```

```
218880_at is significant probeset, with pvalue=0.002027
218945_at is significant probeset, with pvalue=0.002442
218950_at is significant probeset, with pvalue=0.000011
219049_at is significant probeset, with pvalue=0.000514
219060_at is significant probeset, with pvalue=0.000605
219117_s_at is significant probeset, with pvalue=0.001555
219120_at is significant probeset, with pvalue=0.000970
219123_at is significant probeset, with pvalue=0.000175
219244_s_at is significant probeset, with pvalue=0.002209
219299_at is significant probeset, with pvalue=0.001494
219405_at is significant probeset, with pvalue=0.000636
219410_at is significant probeset, with pvalue=0.000435
219450_at is significant probeset, with pvalue=0.000620
219489_s_at is significant probeset, with pvalue=0.002967
219531_at is significant probeset, with pvalue=0.000317
219534_x_at is significant probeset, with pvalue=0.000930
219563_at is significant probeset, with pvalue=0.001660
219641_at is significant probeset, with pvalue=0.000198
219743_at is significant probeset, with pvalue=0.002119
219765_at is significant probeset, with pvalue=0.000845
219785_s_at is significant probeset, with pvalue=0.002956
219928_s_at is significant probeset, with pvalue=0.000014
219944_at is significant probeset, with pvalue=0.000130
219966_x_at is significant probeset, with pvalue=0.000080
219985_at is significant probeset, with pvalue=0.002990
219995_s_at is significant probeset, with pvalue=0.002845
220003_at is significant probeset, with pvalue=0.000216
220177_s_at is significant probeset, with pvalue=0.000491
220197_at is significant probeset, with pvalue=0.000019
220354_at is significant probeset, with pvalue=0.002719
220610_s_at is significant probeset, with pvalue=0.000160
220907_at is significant probeset, with pvalue=0.002133
220935_s_at is significant probeset, with pvalue=0.002283
221016_s_at is significant probeset, with pvalue=0.000154
221024_s_at is significant probeset, with pvalue=0.002248
221029_s_at is significant probeset, with pvalue=0.001750
221096_s_at is significant probeset, with pvalue=0.000849
221532_s_at is significant probeset, with pvalue=0.000521
221538_s_at is significant probeset, with pvalue=0.000734
221567_at is significant probeset, with pvalue=0.001012
221575_at is significant probeset, with pvalue=0.000205
221619_s_at is significant probeset, with pvalue=0.001459
221675_s_at is significant probeset, with pvalue=0.000214
221712_s_at is significant probeset, with pvalue=0.000583
221741_s_at is significant probeset, with pvalue=0.001407
221867_at is significant probeset, with pvalue=0.000277
221909_at is significant probeset, with pvalue=0.001409
221979_at is significant probeset, with pvalue=0.001666
221988_at is significant probeset, with pvalue=0.000354
```

```
222072_at is significant probeset, with pvalue=0.000410
222088_s_at is significant probeset, with pvalue=0.000357
222113_s_at is significant probeset, with pvalue=0.002484
222288_at is significant probeset, with pvalue=0.001368
222455_s_at is significant probeset, with pvalue=0.000254
222537_s_at is significant probeset, with pvalue=0.000424
222561_at is significant probeset, with pvalue=0.000433
222732_at is significant probeset, with pvalue=0.000817
222757_s_at is significant probeset, with pvalue=0.001929
222838_at is significant probeset, with pvalue=0.002281
222912_at is significant probeset, with pvalue=0.001078
223040_at is significant probeset, with pvalue=0.001601
223120_at is significant probeset, with pvalue=0.000541
223121_s_at is significant probeset, with pvalue=0.001136
223122_s_at is significant probeset, with pvalue=0.000980
223130_s_at is significant probeset, with pvalue=0.002445
223168_at is significant probeset, with pvalue=0.001459
223169_s_at is significant probeset, with pvalue=0.001727
223244_s_at is significant probeset, with pvalue=0.000521
223276_at is significant probeset, with pvalue=0.002001
223378_at is significant probeset, with pvalue=0.001605
223424_s_at is significant probeset, with pvalue=0.001909
223442_at is significant probeset, with pvalue=0.002407
223593_at is significant probeset, with pvalue=0.002351
223639_s_at is significant probeset, with pvalue=0.001485
223658_at is significant probeset, with pvalue=0.000256
223721_s_at is significant probeset, with pvalue=0.002777
223792_at is significant probeset, with pvalue=0.001582
223821_s_at is significant probeset, with pvalue=0.001814
223822_at is significant probeset, with pvalue=0.000660
224188_s_at is significant probeset, with pvalue=0.002488
224279_s_at is significant probeset, with pvalue=0.000023
224325_at is significant probeset, with pvalue=0.000045
224570_s_at is significant probeset, with pvalue=0.000445
224690_at is significant probeset, with pvalue=0.000844
224693_at is significant probeset, with pvalue=0.002026
224772_at is significant probeset, with pvalue=0.002052
225005_at is significant probeset, with pvalue=0.000943
225016_at is significant probeset, with pvalue=0.000106
225030_at is significant probeset, with pvalue=0.001664
225105_at is significant probeset, with pvalue=0.000386
225116_at is significant probeset, with pvalue=0.000419
225117_at is significant probeset, with pvalue=0.000515
225252_at is significant probeset, with pvalue=0.000257
225296_at is significant probeset, with pvalue=0.002646
225311_at is significant probeset, with pvalue=0.002322
225337_at is significant probeset, with pvalue=0.000106
225357_s_at is significant probeset, with pvalue=0.001234
225395_s_at is significant probeset, with pvalue=0.000235
```

```
225402_at is significant probeset, with pvalue=0.000064
225609_at is significant probeset, with pvalue=0.001242
225637_at is significant probeset, with pvalue=0.002003
225642_at is significant probeset, with pvalue=0.001518
225651_at is significant probeset, with pvalue=0.002552
225703_at is significant probeset, with pvalue=0.002305
225704_at is significant probeset, with pvalue=0.000020
225843_at is significant probeset, with pvalue=0.000708
225844_at is significant probeset, with pvalue=0.001607
225851_at is significant probeset, with pvalue=0.002945
225909_at is significant probeset, with pvalue=0.000453
225962_at is significant probeset, with pvalue=0.000510
226013_at is significant probeset, with pvalue=0.002511
226032_at is significant probeset, with pvalue=0.000334
226076_s_at is significant probeset, with pvalue=0.001058
226093_at is significant probeset, with pvalue=0.000067
226116_at is significant probeset, with pvalue=0.000760
226139_at is significant probeset, with pvalue=0.002316
226176_s_at is significant probeset, with pvalue=0.000904
226194_at is significant probeset, with pvalue=0.000216
226213_at is significant probeset, with pvalue=0.002488
226224_at is significant probeset, with pvalue=0.002804
226226_at is significant probeset, with pvalue=0.001320
226336_at is significant probeset, with pvalue=0.002163
226443_at is significant probeset, with pvalue=0.000281
226509_at is significant probeset, with pvalue=0.000084
226572_at is significant probeset, with pvalue=0.001758
226780_s_at is significant probeset, with pvalue=0.002010
226781_at is significant probeset, with pvalue=0.000107
226787_at is significant probeset, with pvalue=0.001506
226955_at is significant probeset, with pvalue=0.001899
226976_at is significant probeset, with pvalue=0.002668
227084_at is significant probeset, with pvalue=0.000335
227085_at is significant probeset, with pvalue=0.002997
227155_at is significant probeset, with pvalue=0.001201
227168_at is significant probeset, with pvalue=0.002330
227184_at is significant probeset, with pvalue=0.001430
227197_at is significant probeset, with pvalue=0.001079
227328_at is significant probeset, with pvalue=0.001114
227376_at is significant probeset, with pvalue=0.000898
227388_at is significant probeset, with pvalue=0.000671
227399_at is significant probeset, with pvalue=0.001048
227405_s_at is significant probeset, with pvalue=0.000388
227446_s_at is significant probeset, with pvalue=0.001419
227475_at is significant probeset, with pvalue=0.000297
227480_at is significant probeset, with pvalue=0.001785
227515_at is significant probeset, with pvalue=0.000186
227516_at is significant probeset, with pvalue=0.000510
227522_at is significant probeset, with pvalue=0.001448
```

```
227527_at is significant probeset, with pvalue=0.002684
227538_at is significant probeset, with pvalue=0.002607
227558_at is significant probeset, with pvalue=0.000474
227572_at is significant probeset, with pvalue=0.000718
227615_at is significant probeset, with pvalue=0.000151
227702_at is significant probeset, with pvalue=0.001213
227743_at is significant probeset, with pvalue=0.002013
227852_at is significant probeset, with pvalue=0.001690
227898_s_at is significant probeset, with pvalue=0.001096
227909_at is significant probeset, with pvalue=0.002457
227923_at is significant probeset, with pvalue=0.001327
227925_at is significant probeset, with pvalue=0.002673
228055_at is significant probeset, with pvalue=0.000070
228093_at is significant probeset, with pvalue=0.000848
228292_at is significant probeset, with pvalue=0.002537
228346_at is significant probeset, with pvalue=0.002273
228442_at is significant probeset, with pvalue=0.002190
228461_at is significant probeset, with pvalue=0.001090
228490_at is significant probeset, with pvalue=0.000842
228664_at is significant probeset, with pvalue=0.002061
228790_at is significant probeset, with pvalue=0.002260
228809_at is significant probeset, with pvalue=0.001525
228811_at is significant probeset, with pvalue=0.000588
228854_at is significant probeset, with pvalue=0.000527
228899_at is significant probeset, with pvalue=0.000555
228944_at is significant probeset, with pvalue=0.002428
228967_at is significant probeset, with pvalue=0.001996
229086_at is significant probeset, with pvalue=0.001169
229101_at is significant probeset, with pvalue=0.000434
229106_at is significant probeset, with pvalue=0.001998
229157_at is significant probeset, with pvalue=0.000034
229158_at is significant probeset, with pvalue=0.000093
229175_at is significant probeset, with pvalue=0.001200
229213_at is significant probeset, with pvalue=0.001824
229265_at is significant probeset, with pvalue=0.001067
229281_at is significant probeset, with pvalue=0.001660
229302_at is significant probeset, with pvalue=0.001173
229309_at is significant probeset, with pvalue=0.000480
229356_x_at is significant probeset, with pvalue=0.000386
229377_at is significant probeset, with pvalue=0.000244
229407_at is significant probeset, with pvalue=0.001937
229410_at is significant probeset, with pvalue=0.001455
229537_at is significant probeset, with pvalue=0.000488
229555_at is significant probeset, with pvalue=0.001427
229566_at is significant probeset, with pvalue=0.000375
229606_at is significant probeset, with pvalue=0.002824
229742_at is significant probeset, with pvalue=0.001818
229852_at is significant probeset, with pvalue=0.001469
229964_at is significant probeset, with pvalue=0.000064
```

```
229977_at is significant probeset, with pvalue=0.002312
230054_at is significant probeset, with pvalue=0.000903
230100_x_at is significant probeset, with pvalue=0.000265
230130_at is significant probeset, with pvalue=0.002333
230136_at is significant probeset, with pvalue=0.002122
230311_s_at is significant probeset, with pvalue=0.000277
230433_at is significant probeset, with pvalue=0.000001
230472_at is significant probeset, with pvalue=0.002175
230747_s_at is significant probeset, with pvalue=0.000056
230776_at is significant probeset, with pvalue=0.000492
230782_at is significant probeset, with pvalue=0.000032
230849_at is significant probeset, with pvalue=0.001047
230857_s_at is significant probeset, with pvalue=0.002299
230888_at is significant probeset, with pvalue=0.002385
230999_at is significant probeset, with pvalue=0.001741
231232_at is significant probeset, with pvalue=0.002204
231362_at is significant probeset, with pvalue=0.002042
231379_at is significant probeset, with pvalue=0.001154
231487_at is significant probeset, with pvalue=0.002046
231779_at is significant probeset, with pvalue=0.002940
231800_s_at is significant probeset, with pvalue=0.000275
231815_at is significant probeset, with pvalue=0.002426
231835_at is significant probeset, with pvalue=0.000670
231845_at is significant probeset, with pvalue=0.001620
231849_at is significant probeset, with pvalue=0.002116
231901_at is significant probeset, with pvalue=0.000024
231907_at is significant probeset, with pvalue=0.000505
231928_at is significant probeset, with pvalue=0.002965
231952_at is significant probeset, with pvalue=0.002708
232059_at is significant probeset, with pvalue=0.000626
232303_at is significant probeset, with pvalue=0.001209
232704_s_at is significant probeset, with pvalue=0.000477
232766_at is significant probeset, with pvalue=0.000453
232891_at is significant probeset, with pvalue=0.001499
233498_at is significant probeset, with pvalue=0.002007
233565_s_at is significant probeset, with pvalue=0.001238
233852_at is significant probeset, with pvalue=0.001698
234073_at is significant probeset, with pvalue=0.002046
234148_at is significant probeset, with pvalue=0.001688
234317_s_at is significant probeset, with pvalue=0.002583
234329_at is significant probeset, with pvalue=0.002971
235085_at is significant probeset, with pvalue=0.000325
235214_at is significant probeset, with pvalue=0.002924
235233_s_at is significant probeset, with pvalue=0.002906
235350_at is significant probeset, with pvalue=0.000664
235362_at is significant probeset, with pvalue=0.001541
235459_at is significant probeset, with pvalue=0.001690
235533_at is significant probeset, with pvalue=0.001370
235647_at is significant probeset, with pvalue=0.002374
```

```
235691_at is significant probeset, with pvalue=0.001012
235727_at is significant probeset, with pvalue=0.001932
235751_s_at is significant probeset, with pvalue=0.001025
235793_at is significant probeset, with pvalue=0.002785
235804_at is significant probeset, with pvalue=0.001372
235837_at is significant probeset, with pvalue=0.000829
235948_at is significant probeset, with pvalue=0.000492
236124_at is significant probeset, with pvalue=0.000468
236132_at is significant probeset, with pvalue=0.000254
236465_at is significant probeset, with pvalue=0.000337
236656_s_at is significant probeset, with pvalue=0.000046
236668_at is significant probeset, with pvalue=0.001720
236833_at is significant probeset, with pvalue=0.001777
237330_at is significant probeset, with pvalue=0.001653
237351_at is significant probeset, with pvalue=0.000018
237721_s_at is significant probeset, with pvalue=0.002886
238369_s_at is significant probeset, with pvalue=0.002688
238425_at is significant probeset, with pvalue=0.002581
238755_at is significant probeset, with pvalue=0.000060
238999_at is significant probeset, with pvalue=0.001372
239021_at is significant probeset, with pvalue=0.001830
239093_at is significant probeset, with pvalue=0.000212
239142_at is significant probeset, with pvalue=0.002077
239205_s_at is significant probeset, with pvalue=0.000703
239207_at is significant probeset, with pvalue=0.002516
239229_at is significant probeset, with pvalue=0.000613
239283_at is significant probeset, with pvalue=0.001007
239433_at is significant probeset, with pvalue=0.000609
239436_at is significant probeset, with pvalue=0.002503
239595_at is significant probeset, with pvalue=0.000146
240155_x_at is significant probeset, with pvalue=0.001197
240382_at is significant probeset, with pvalue=0.002267
240454_at is significant probeset, with pvalue=0.000707
240555_at is significant probeset, with pvalue=0.000150
240699_at is significant probeset, with pvalue=0.000091
240788_at is significant probeset, with pvalue=0.000314
240867_at is significant probeset, with pvalue=0.001654
240869_at is significant probeset, with pvalue=0.000181
240899_at is significant probeset, with pvalue=0.000333
241233_x_at is significant probeset, with pvalue=0.000951
241418_at is significant probeset, with pvalue=0.000725
241764_at is significant probeset, with pvalue=0.001327
241877_at is significant probeset, with pvalue=0.000009
241890_at is significant probeset, with pvalue=0.000244
241950_at is significant probeset, with pvalue=0.001224
241990_at is significant probeset, with pvalue=0.002616
242065_x_at is significant probeset, with pvalue=0.002142
242271_at is significant probeset, with pvalue=0.002500
242452_at is significant probeset, with pvalue=0.002250
```

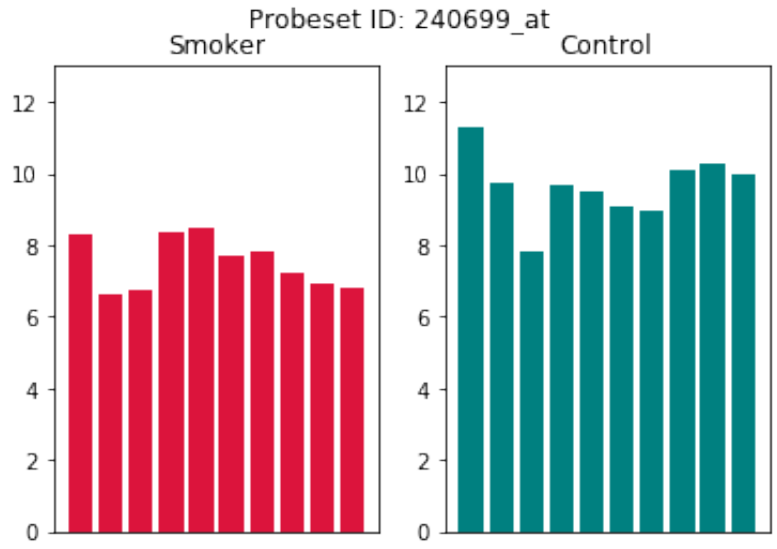
```
242478_at is significant probeset, with pvalue=0.000508
243594_x_at is significant probeset, with pvalue=0.000634
243671_at is significant probeset, with pvalue=0.001411
243934_at is significant probeset, with pvalue=0.002834
244362_at is significant probeset, with pvalue=0.001491
244519_at is significant probeset, with pvalue=0.001042
244589_at is significant probeset, with pvalue=0.002393
244654_at is significant probeset, with pvalue=0.001966
244677_at is significant probeset, with pvalue=0.001759
31845_at is significant probeset, with pvalue=0.001272
36129_at is significant probeset, with pvalue=0.002983
36612_at is significant probeset, with pvalue=0.001874
37424_at is significant probeset, with pvalue=0.002685
40284_at is significant probeset, with pvalue=0.000008
41577_at is significant probeset, with pvalue=0.000194
42361_g_at is significant probeset, with pvalue=0.000157
47560_at is significant probeset, with pvalue=0.002922
50221_at is significant probeset, with pvalue=0.000382
52731_at is significant probeset, with pvalue=0.001303
53720_at is significant probeset, with pvalue=0.002935
55081_at is significant probeset, with pvalue=0.000410
55872_at is significant probeset, with pvalue=0.000368
65630_at is significant probeset, with pvalue=0.001782
823_at is significant probeset, with pvalue=0.000334
91617_at is significant probeset, with pvalue=0.000163
642
```

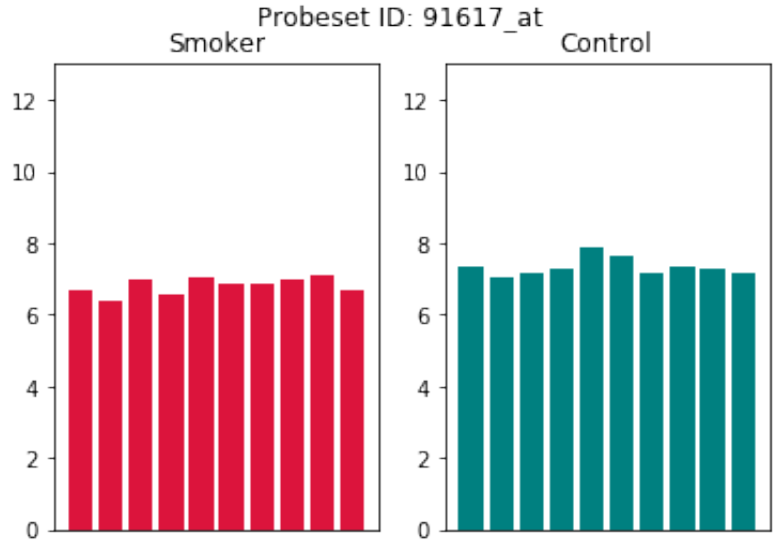
(d) T-test Vistualization

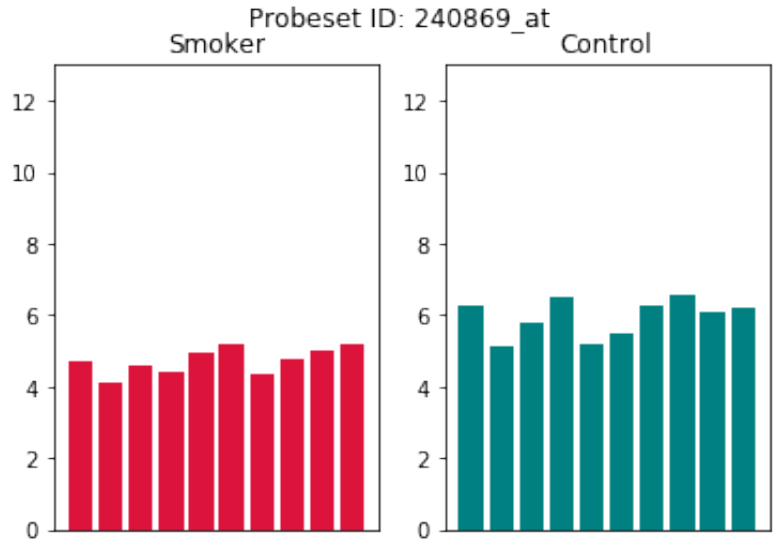
select Gene: '40699_at, 91617_at, 240869_at'

- Bar Plot
- Scatter Plot

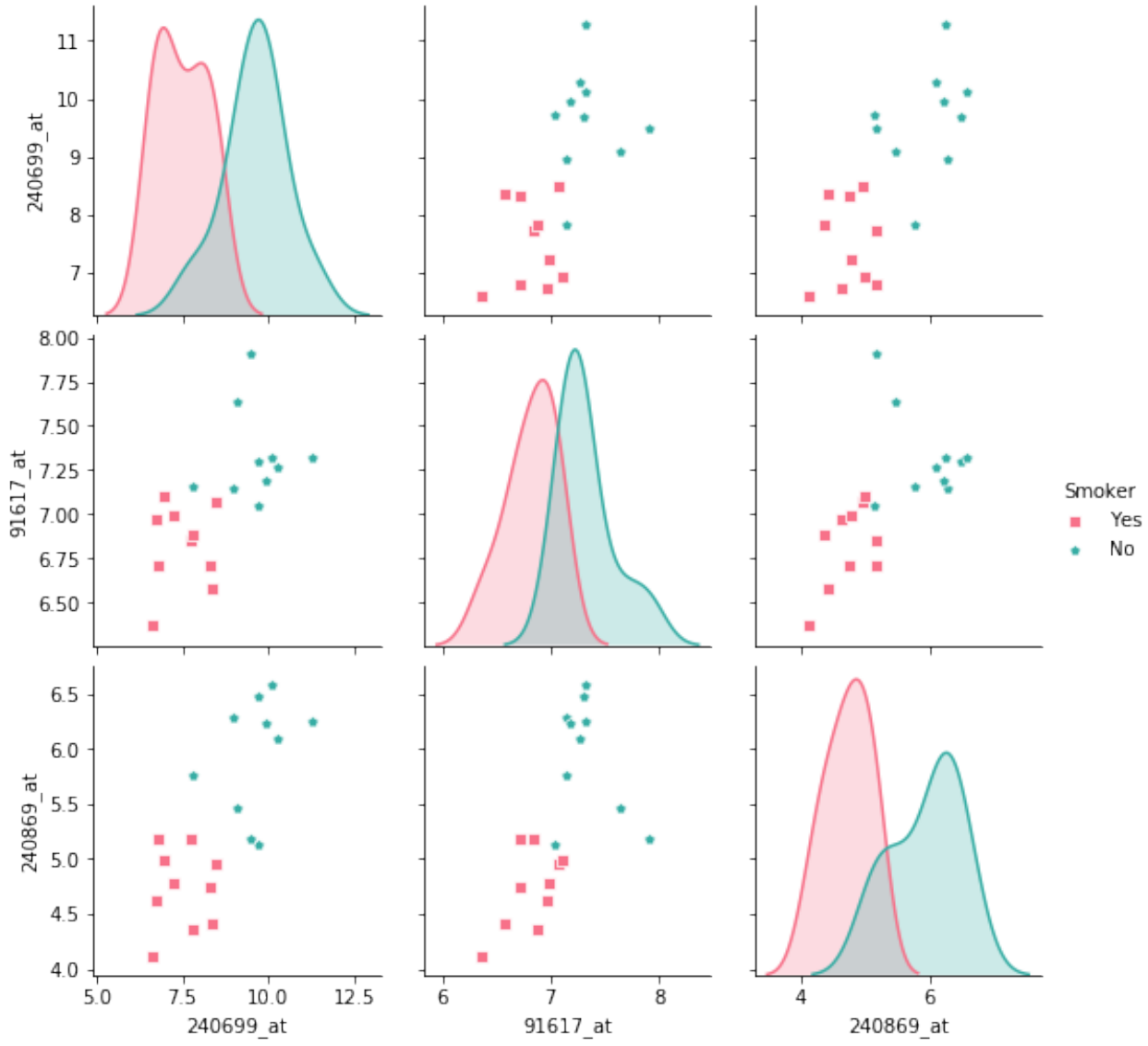
```
select_w_list= ['240699_at', '91617_at', '240869_at']
smoker_ttest = smoker_filt[non_smoker_list + smoker_list].loc[select_w_list]
for ii in smoker_ttest.index:
    fig = plt.figure()
    ax1 = fig.add_subplot(121)
    ax1.bar(smoker_list, smoker_ttest[smoker_list].loc[ii], color="#dc143c")
    plt.xticks(''); plt.title("Smoker"); plt.ylim(0, 13)
    ax2 = fig.add_subplot(122)
    ax2.bar(smoker_list, smoker_ttest[non_smoker_list].loc[ii],
color="#008080")
    plt.xticks(''); plt.title("Control"); plt.ylim(0, 13)
    fig.suptitle("Probeset ID: %s" % ii)
```


Probeset ID: 240699_at
Smoker
Control
0
2
4
6
8
10
12


Probeset ID: 91617_at
Smoker
Control
0
2
4
6
8
10
12


Probeset ID: 240869_at
Smoker
Control
0
2
4
6
8
10
12

```
# Scatter plots combining the probesets
df_scatter_s = smoker_ttest[smoker_list].T
df_scatter_s['Smoker'] = 'Yes'
df_scatter_c = smoker_ttest[non_smoker_list].T
df_scatter_c['Smoker'] = 'No'
df_scatter = pd.concat([df_scatter_s, df_scatter_c], axis=0)
# print(df_scatter)
# print(df_scatter.index)
# Pairplot
pp = sns.pairplot(df_scatter, hue='Smoker', diag_kind='auto', markers=['s',
'p'], palette="husl")
```



Question 3 Quiz: Comment the results based on your observations

Anwser

Firstly, up/down regulation in 6 selected genes: observing the bar plot, reader can notice the expression ratio in test(smoke) and control(non-smoke) samples are up/down regulated. compare the reference group for each samples:

- 1552307_a_at: down regulation
- 1552497_a_at: up regulation
- 1552834_at: up regulation
- 240699_at: down regulation
- 91617_at: down regulation
- 240869_at: down regulation

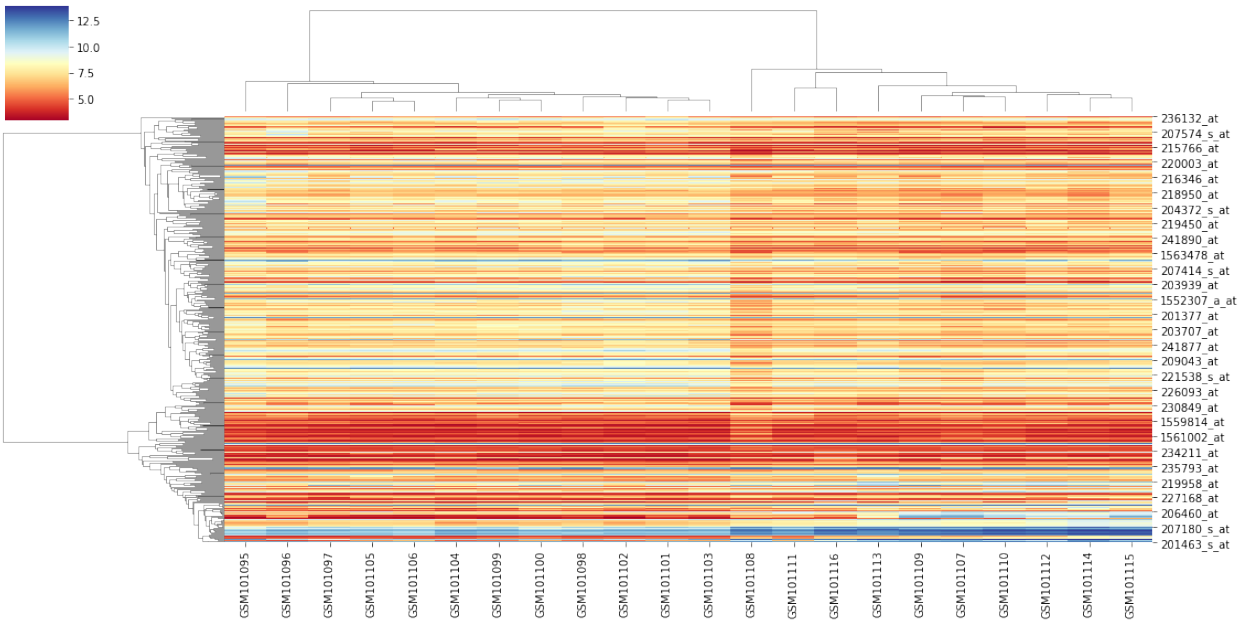
Secondly, for a specific gene to compare, though basically exists up or down regulation, the level of up or down regulation is different in different samples. For example, take reference ID: 1552497_a_at as an example, the second sample in the smoker group is relatively in high expression ratio compared to others in the same group.

Thirdly, the scatter plot indicates for the correlations/similarity between two selected genes: some are not correlated as the points can be divided into two clusters easily; some are correlated as the points are mixed together. The "correlation" might be the evidence for the genes control the same functional categories, which would be mentioned in the part (h).

(g) Question 4: Hierarchical Clustering and Dendrogram

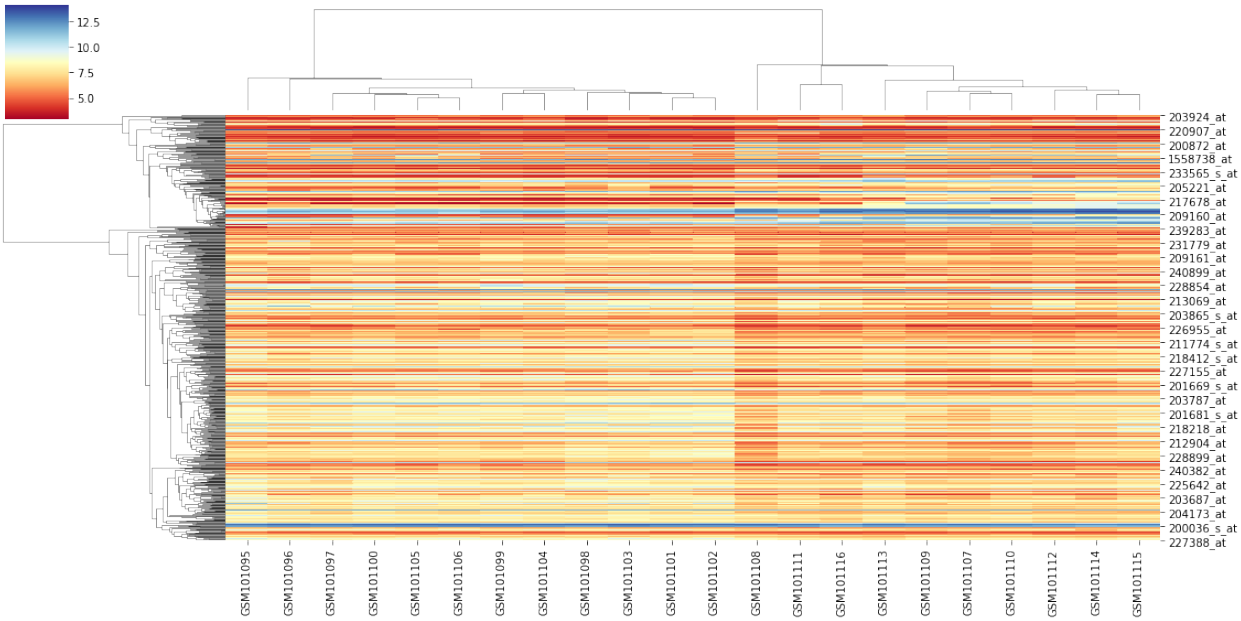
- Wilcoxon Signed-Rank Test
- T-test

```
# Wilcoxon Signed-Rank Test
w_target_df = pd.DataFrame(smoker_filt, index=significant_inds_wilcox)
w_hc = sns.clustermap(w_target_df, method='average', metric='correlation',
cmap='RdYlBu', figsize=(16, 8))
```



```
# T-test
t_target_df = pd.DataFrame(smoker_filt, index=significant_inds_ttest)
t_hc = sns.clustermap(t_target_df, method='average', metric='correlation',
cmap='RdYlBu', figsize=(16, 8))
```

```
/Library/Frameworks/Python.framework/Versions/3.7/lib/python3.7/site-
packages/seaborn/matrix.py:624: UserWarning: Clustering large matrix with
scipy. Installing `fastcluster` may give better performance.
  warnings.warn(msg)
```



Question 4 Quiz: Give a brief interpretation for the results.

Anwser

- From Gene: most of the selected genes are up regulated or normal regulated, only an extremely part is down regulated (e.g. 207180_s_at in Wilcoxon Signed-Rank Test).
- From Sample: it can be observed that the heatmap is splited into two "different" parts, which implies different expression patterns for control and test sample sets; in the same set(i.e. control set or test set), the expression patterns are quite similar.
- From Dendrogram:

1. the relationships among different samples in one set(i.e. control set or test set) are quite random, however, the relationship among different sets is absolute the most estranged one.
2. the genes have most different expression pattern, have the most estranged relation; vice versa.
3. the genes have close relationship or be clustered into the same group should be highlighted,

since they might control the similar/same functions.

- Although two test selected different differential expressed genes, for the same genes in two tests, the absolute expression value might be different, but the expression pattern rule remains similar.

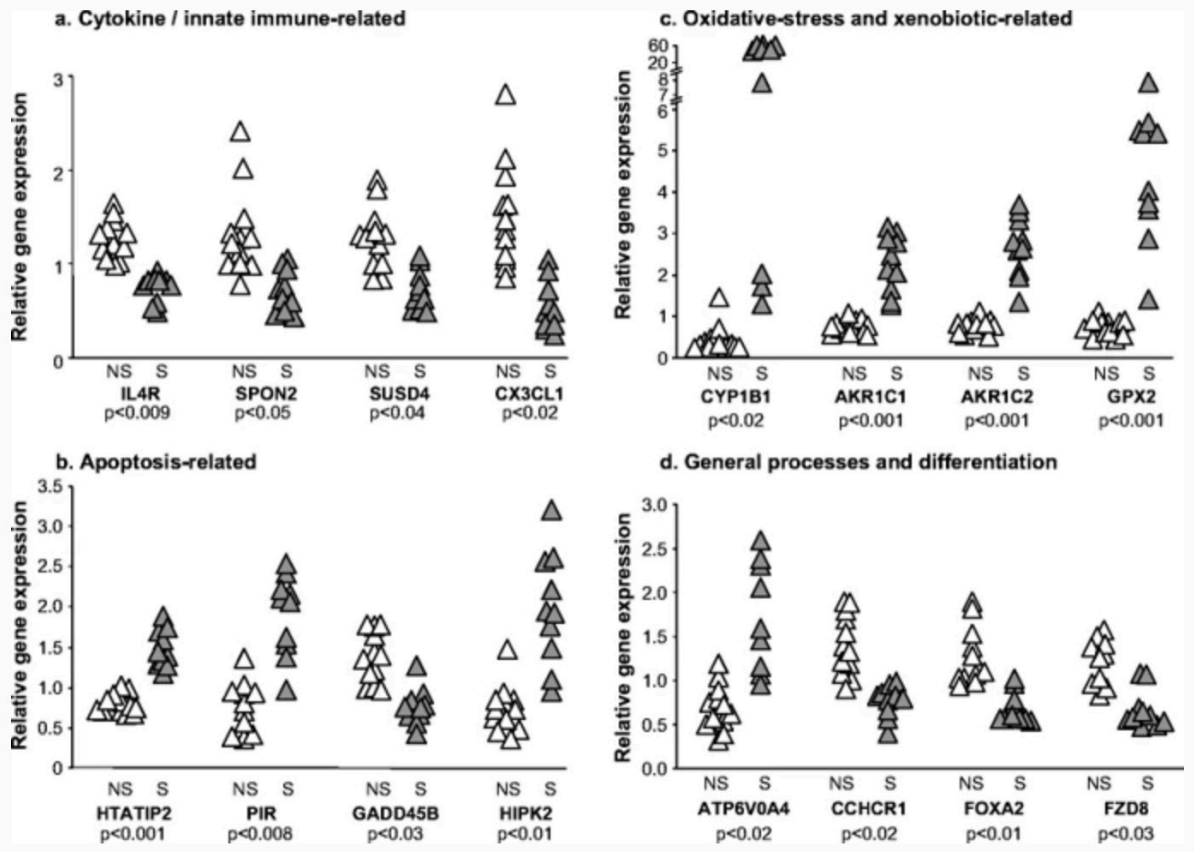
(h) Additional Part: Further Comments

Introduction for Dataset

- This experiment aims to understand how smoking modifies small airway structure and function. The dataset provided in the assignemnt is the second dataset in this project. - Although the smokers were phenotypically normal, microarray analysis of gene expression of the small airway epithelium of the smokers compared to the nonsmokers demonstrated up- and down regulation of genes in multiple categories relevant to the pathogenesis of chronic obstructive lung disease (COPD), including genes coding for cytokines/innate immunity, apoptosis, mucin, response to oxidants and xenobiotics, and general cellular processes.

Optimize Method for The Analysis of Assignment 4

- Apply different p-value thresholds for different genes
  The Wilcoxon Signed-Rank test and Paired Sample T-Test here we applied only set one p-value for all the genes, however, the differential level for different genes diverse a lot. For example, the interleukin-4 (IL4) receptor gene with $p < 0.002$, the chemokine (C-X3-C motif) ligand 1 with $p < 0.02$, the spondin 2 with $p < 0.04$.



- Check the differential differences between the samples for constant variables.
  If we want to compare different samples, we need to make sure the differences brought by sample itself will not cause large impact, that is the pvalue for factors like age, gender and etc. should reject the test: no differences in age with $p > 0.2$, sex with $p > 0.6$, or race with $p > 0.7$ among the smokers and nonsmokers.
- Cluster the genes according to functional category.
  It would be more signifcant if we cluster the gene to identify the gene function.