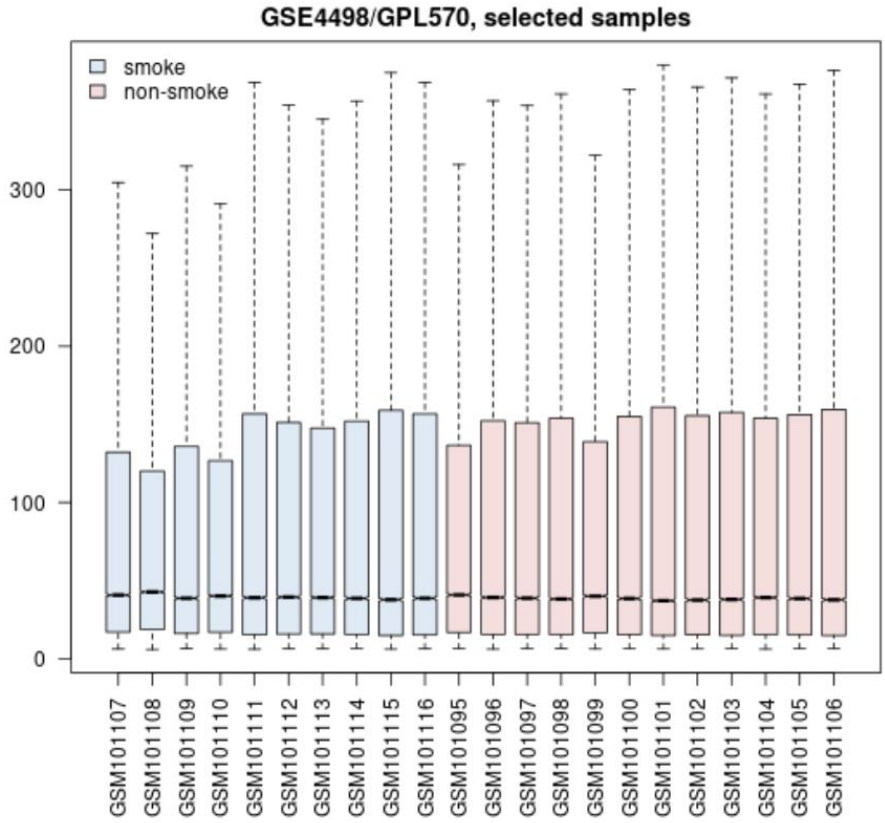
BIM3001 homewoke4

117020290

**Q1**

**Data pretreatment**




The dataset we use is from GEO data base. Expression level for each sample can be seen from website. From this figure, we can know that normalization has been done, since the mean level of expression of all sample is at the same level. You can also perform log2 transformation to each sample. Calculate mean of each sample, divide all values by its sample mean and take log2 transformation. This is used to remove the bias of extremely large value. In this project, I select DEG use origin expression value from dataset and use both origin value and log2 transfored data when clustering.

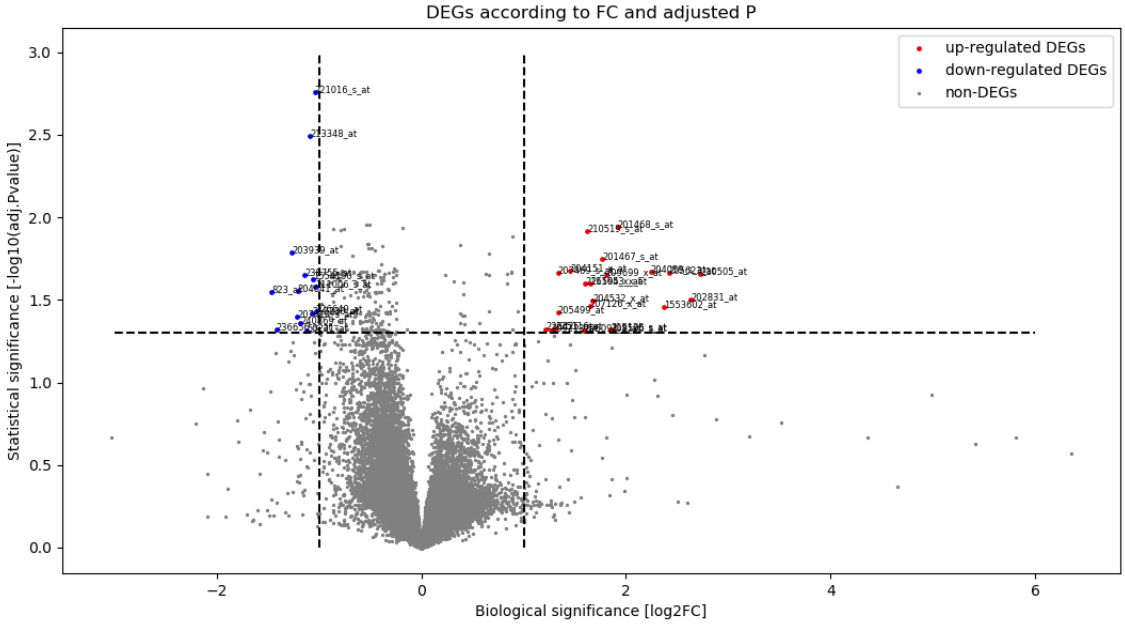
**Differential expressed genes selection**

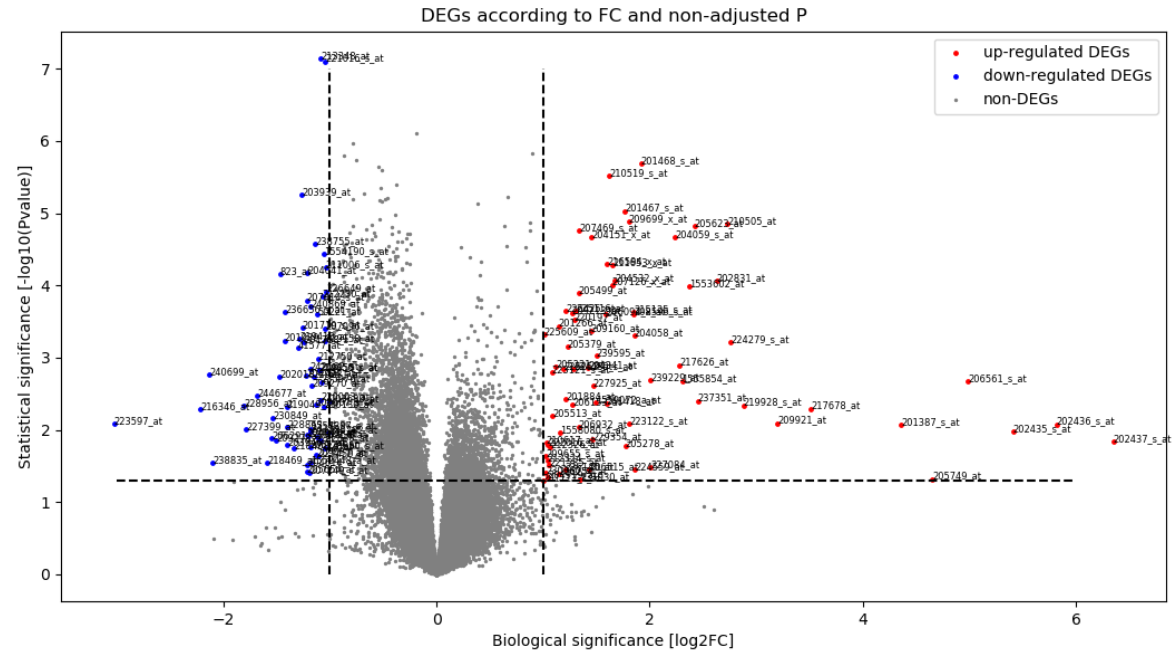
In this question, I use both biological and statistical methods to select differential expressed genes (DEGs). I use log2 fold change to select genes between -1 and 1. I use t-test and perform BH adjustment to P value, select significantly differential genes with adjust p value smaller than 0.05. Program is enclosed in another python file.

You can see this figure intuitionally: 36 differential expression genes are selected out on both up corner of the figure. Genes in the red circle are up regulated genes and genes in the blue circle are down regulated genes.

In order to compare different statistical methods, I also select differential expression genes according to non-adjust P values. Since the criteria becomes looser, 138 DEGs

are selected out. Blue points on left up corner of the figure represent down regulated DEGs and red points on right up corner of the figure represent up regulated DEGs.




**Functional enrichment analysis results based on DEGs**

After loading DEG gene list to DAVID and download functional enrichment results data, I selected several functions or pathway with FDR adjusted P value lower than 0.05 or functions with more DEGs involved.

**Based on DEGs selected from BH adjust P value and FC**

Compared to databases such as UP_KEYWORDS, GOTERM_BP_DIRECT,

KEGG_PATHWAY, and GOTERM_MF_DIRECT, These DEGs are related to oxidoreductase, oxidoreductase activity, NADP, oxidation-reduction process, metabolism of xenobiotics by cytochrome P450, and cytoplasm.

| | | |
|---|---|---|
| UP_KEYWORDS | oxidoreductase | Enzyme that catalyzes the oxidation of one compound with the reduction of another. |
| GOTERM_BP_DIRECT | Oxidation-reduction process | A metabolic process that results in the removal or addition of one or more electrons to or from a substance, with or without the concomitant removal or addition of a proton or protons |
| KEGG_PATHWAY | metabolism of xenobiotics by cytochrome P450 | Relate to several oxidation-reduction enzymes such as alcohol dehydrogenase, carbonyl reductase, aldo-keto reductase, UDP gluronosyltrasferase |
| UP_KEYWORDS | NADP | NADP serves as an electron carrier by being alternately oxidized (NADP+) and reduced (NADPH) |
| UP_KEYWORDS | cytoplasm | three-dimensional, jelly-like lattice, and it interconnects and supports the other solid structures. |

These results suggest that smoke affect oxidation-reduction process by affecting enzyme activity and substrates. It may also affect function or component of cytoplasm.

**Based on DEGs selected from non-adjust P value and FC**

All results from previous section are found significantly important this time. What's more, lung cancer is found to have a correlation with these genes. Smoke may affect people's health through affecting their oxidation-reduction metabolism. Detailed information is attached in files.

**Q2**

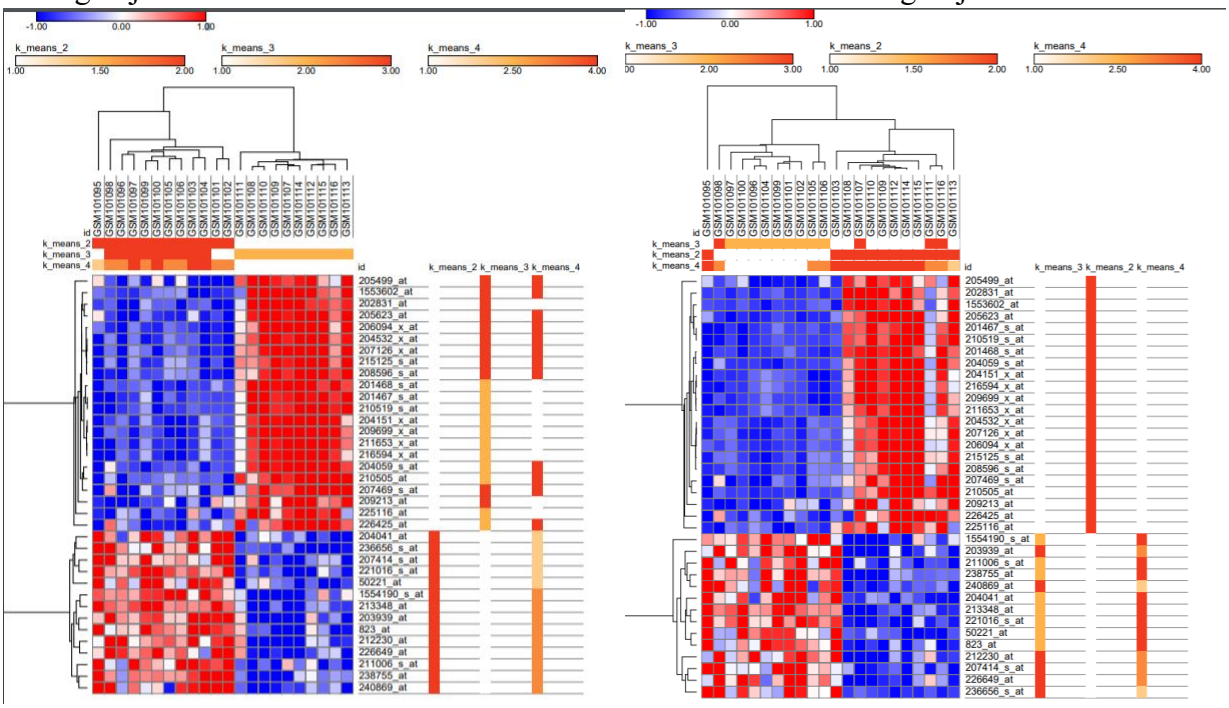
**Clustering results**

In this section, I separate samples and genes into different number of clusters based on hierarchical and non-hierarchical (K-means) methods. The tool I used to do this is

Morpheus, and its website is: https://software.broadinstitute.org/morpheus/. When doing the hierarchical clustering, Pearson correlation with average linkage method is used to calculate distance between vectors. When doing k-means clustering, Person correlation is also used to calculate distance between vectors.

To test whether log2 pre-treatment of data will enhance the clustering results. I select data both from log2 treatment datasheet and original datasheet. Log2 pretreatment details(I have mentioned at the beginning of this report) are: calculate mean expression value of each sample, divide each value in the sample by the average level, take log2 transformation of the ratio and output new datasheet. This method is supported to decrease bad effects from extremely large value to the results, which may lead to bias when calculating distances.

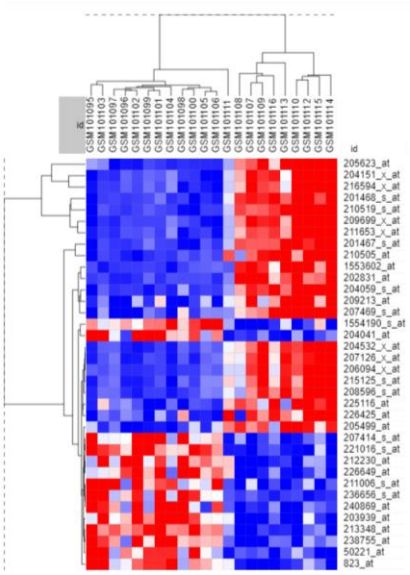
Log-adj nonlog-adj



**Comparation of clustering results of data with log2 pretreatment and without log2 pretreatment**

According to the previous figure, data with log 2 transformation have been clustered perfectly to smoke samples and non-smoke samples. However, there are problems with clustering results from data without log2 pretreatment.

Pearson correlation method to calculate distance may not be suitable for this condition, thus I try Euclidean to calculated distance and do clustering to DEG data without pretreatment again as following. You can see that, although it seems better than using Pearson correlation methods, one sample ‘GSM101111’, which is supposed to be classified into smoke samples, is clustered into non-smoke samples.

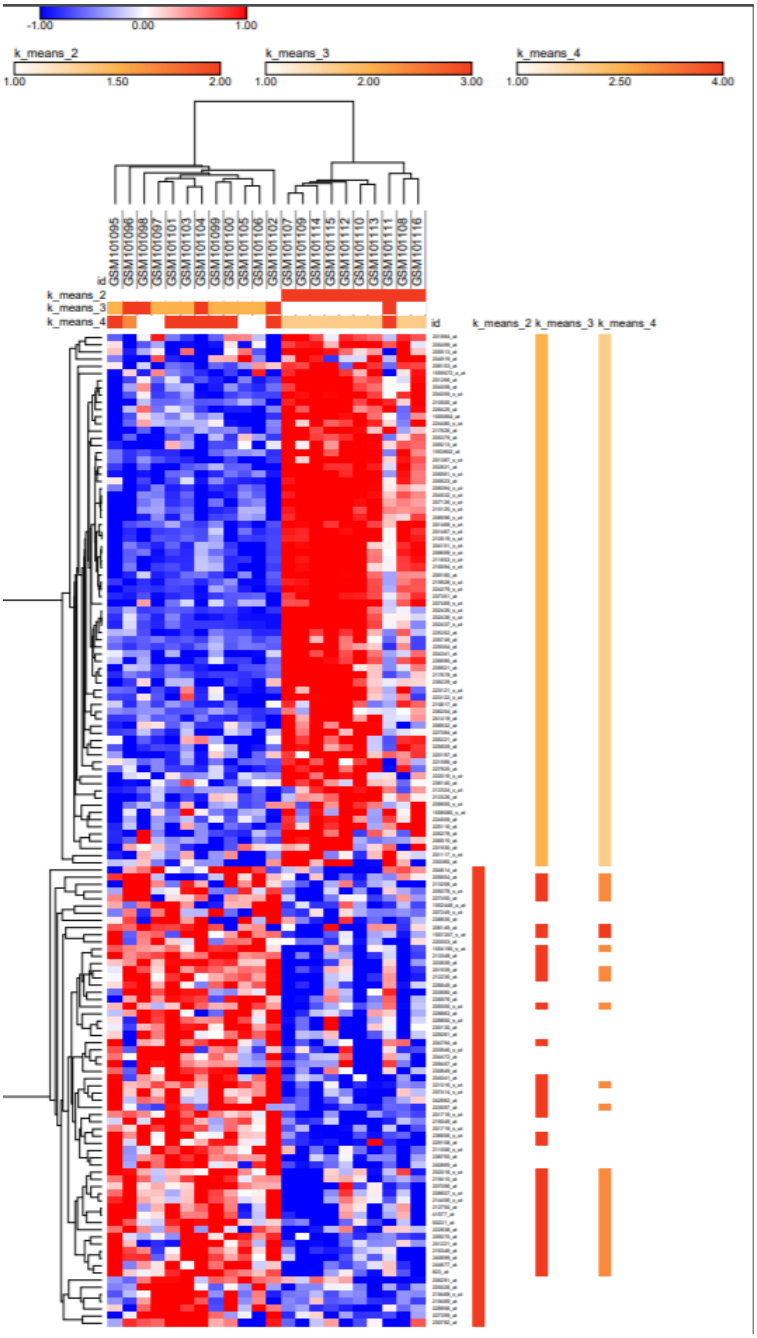



Thus, we can conclude that log2 transformation pretreatment of data can remove bias of extreme value and enhance the clustering performance. We can also conclude from results that there is something different with sample ‘GSM101111’. Its gene expression pattern is not as significant as other samples in smoke cluster. I guess that this people may just start smoke not for a long time or under other conditions that different from other smoke samples. The age of sample ‘GSM10111’ is 37 years old, which is the youngest among smokers, and this may provide explanation of the differences.

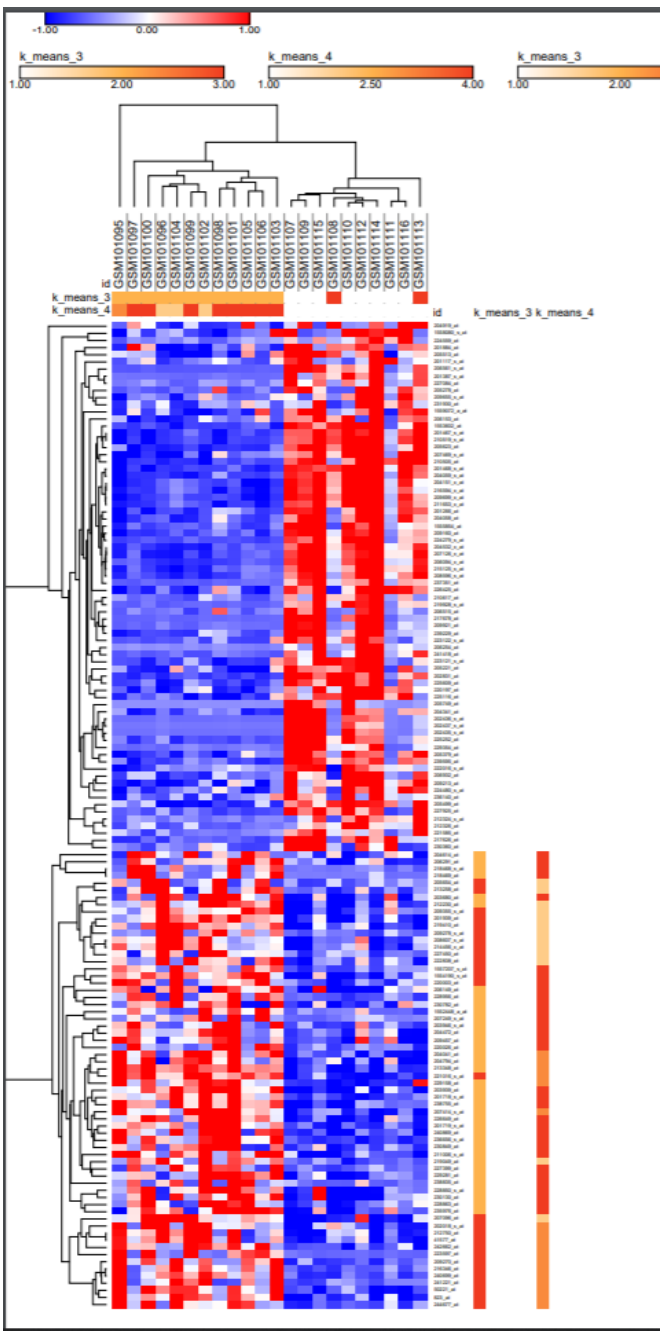
Columns

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| non-smoke | GSM101098 | small airways, non-smoker 003, RMA and MAS | airway epithelial cells obtained by bronchoscopy and brushing | | | | | 37 | F | black | non-smoker |
| non-smoke | GSM101099 | small airways, non-smoker 013, RMA and MAS | airway epithelial cells obtained by bronchoscopy and brushing | 45 | M | hispanic | non-smoker | | | | |
| non-smoke | GSM101100 | small airways, non-smoker 006, RMA and MAS | airway epithelial cells obtained by bronchoscopy and brushing | | | | | 47 | M | black | non-smoker |
| non-smoke | GSM101101 | small airways, non-smoker 021, RMA and MAS | airway epithelial cells obtained by bronchoscopy and brushing | | | | | 38 | M | hispanic | non-smoker |
| non-smoke | GSM101102 | small airways, non-smoker 019, RMA and MAS | airway epithelial cells obtained by bronchoscopy and brushing | | | | | 49 | F | white | non-smoker |
| non-smoke | GSM101103 | small airways, non-smoker 014, RMA and MAS | airway epithelial cells obtained by bronchoscopy and brushing | | | | | 45 | M | white | non-smoker |
| non-smoke | GSM101104 | small airways, non-smoker 008, RMA and MAS | airway epithelial cells obtained by bronchoscopy and brushing | | | | | 36 | M | white | non-smoker |
| non-smoke | GSM101105 | small airways, non-smoker 015, RMA and MAS | airway epithelial cells obtained by bronchoscopy and brushing | | | | | 38 | M | black | non-smoker |
| non-smoke | GSM101106 | small airways, non-smoker 005, RMA and MAS | airway epithelial cells obtained by bronchoscopy and brushing | | | | | 35 | M | black | non-smoker |
| smoke | GSM101107 | small airways, smoker 002, RMA and MAS | airway epithelial cells obtained by bronchoscopy and brushing | | | | | 46 | M | white | smoker, 21 pack-years |
| smoke | GSM101108 | small airways, smoker 003, RMA and MAS | airway epithelial cells obtained by bronchoscopy and brushing | 40 | F | black | smoker, 25 pack-years | | | | |
| smoke | GSM101109 | small airways, smoker 027, RMA and MAS | airway epithelial cells obtained by bronchoscopy and brushing | 44 | M | white | smoker, 45 pack-years | | | | |
| smoke | GSM101110 | small airways, smoker 033, RMA and MAS | airway epithelial cells obtained by bronchoscopy and brushing | 43 | M | white | smoker, 15 pack-years | | | | |
| smoke | GSM101111 | small airways, smoker 001, RMA and MAS | airway epithelial cells obtained by bronchoscopy and brushing | | | | | 37 | F | black | smoker, 23 pack-years |
| smoke | GSM101112 | small airways, smoker 023, RMA and MAS | airway epithelial cells obtained by bronchoscopy and brushing | 41 | M | black | smoker, 20 pack-years | | | | |
| smoke | GSM101113 | small airways, smoker 048, RMA and MAS | airway epithelial cells obtained by bronchoscopy and brushing | | | | | 45 | M | black | smoker, 28 pack-years |
| smoke | GSM101114 | small airways, smoker 041, RMA and MAS | airway epithelial cells obtained by bronchoscopy and brushing | | | | | 48 | M | white | smoker, 20 pack-years |
| smoke | GSM101115 | small airways, smoker 044, RMA and MAS | airway epithelial cells obtained by bronchoscopy and brushing | | | | | 50 | M | white | smoker, 38 pack-years |
| smoke | GSM101116 | small airways, smoker 049, RMA and MAS | airway epithelial cells obtained by bronchoscopy and brushing | | | | | 46 | F | black | smoker, 23 pack-years |

log_Non-adj




non-log_non-adj




Expression pattern of DEGs selected from non-adjusted P value seems not as pure as that selected from adjust P value. But generally speaking, the clustering results based on both genes and samples are good. According to the figure, you can see that sample 'GSM101111', 'GSM101108', 'GSM101106' have different expression pattern compared to other samples in smoke group. Expression level of several genes, which is supposed to be up regulated in smoke group, is down regulated or not such obviously up regulated. This suggest that sample 'GSM101111', 'GSM101108' and 'GSM101106' should be further learned and discussed, especially 'GSM101111'. One guess is that they just start smoke not for a long time. After looking at these three sample, I found that all of them are black women smoker. This may affect their gene expression level.

**Functional enrichment analysis results in David based on clusters**
**Adj-log**
**Up regulated genes function:**

| | |
|---|---|
| UP_KEYWORDS | Oxidoreductase |
| GOTERM_BP_DIRECT | GO:0055114~oxidation-reduction process |
| KEGG_PATHWAY | hsa00980: Metabolism of xenobiotics by cytochrome P450 |
| UP_KEYWORDS | NADP |
| GOTERM_MF_DIRECT | GO:0016655~oxidoreductase activity, acting on NAD(P)H, quinone or similar compound as acceptor |
| UP_SEQ_FEATURE | nucleotide phosphate-binding region: NADP |
| UP_KEYWORDS | Cytoplasm |
| KEGG_PATHWAY | hsa05204: Chemical carcinogenesis |
| GOTERM_MF_DIRECT | GO:0016491~oxidoreductase activity |
| GOTERM_BP_DIRECT | GO:0030855~epithelial cell differentiation |
| GOTERM_BP_DIRECT | GO:0006805~xenobiotic metabolic process |
| KEGG_PATHWAY | hsa00140: Steroid hormone biosynthesis |
| KEGG_PATHWAY | hsa00982: Drug metabolism - cytochrome P450 |
| INTERPRO | IPR016040: NAD(P)-binding domain |

7 of 14 important functions are related to oxidation-reduction process, such as activity of oxidase, NADP, binding domain and so on. Which suggests that smoke may up regulated certain genes that increase oxidation-reduction reaction activity.
Another very important enzyme that these genes associated are Cytochromes. P450 (CYPs) are a family of enzymes containing heme as a cofactor that function as monooxygenases. It relates to oxidation and it also affect drug metabolism. 'Cytochrome P450 (CYP) is a hemeprotein that plays a key role in the metabolism of drugs and other xenobiotics. Drug metabolism is achieved through phase I reactions, phase II reactions, or both. The most common phase I reaction is oxidation, which is catalyzed by the CYP system.' According to chemical carcinogenesis and drug metabolism in KEGG_PATHWAY, Cytochrome P450 is also highly related to several cancers such as Bladder cancer, skin cancer, lung cancer gastric cancer, liver cancer and so on. In conclusion, up regulated genes relate to cytochromes P450. These enzymes affect drug metabolism and participate in oxidoreductase.

These genes also relate to epithelial cell differentiation and steroid hormone biosynthesis.

**Down regulated genes**

| | |
|---|---|
| GOTERM_BP_DIRECT | GO:0030111~regulation of Wnt signaling pathway |
| GOTERM_MF_DIRECT | GO:0005178~integrin binding |
| KEGG_PATHWAY | hsa01100: Metabolic pathways |

This is result from DAVID of down regulated genes. However, P value is not small enough to give significant conclusion.

In conclusion, up regulated genes in smoke sample have major affections on patients. These genes effects oxidation-reduction process, effects drug metabolism and may lead to cancers. Among those cancers, lung cancer and breast cancer are the most possible ones that may occur according to database GAD_DISEASE. Down regulated genes may affect some signaling pathway, but the statistical meaning is not so significant.

**Nonadj_origin**

**Up regulated genes functions**

The functions are statistical significant based on FDR adjust P value.

| Category | Term |
|---|---|
| UP_KEYWORDS | Oxidoreductase |
| GOTERM_BP_DIRECT | GO:0055114~oxidation-reduction process |
| UP_KEYWORDS | NADP |
| KEGG_PATHWAY | hsa00980: Metabolism of xenobiotics by cytochrome P450 |
| GAD_DISEASE | Lung Cancer |
| UP_SEQ_FEATURE | nucleotide phosphate-binding region: NADP |
| INTERPRO | IPR023210: NADP-dependent oxidoreductase domain |
| GOTERM_MF_DIRECT | GO:0016655~oxidoreductase activity, acting on NAD(P)H, quinone or similar compound as acceptor |
| GOTERM_MF_DIRECT | GO:0016491~oxidoreductase activity |
| GOTERM_BP_DIRECT | GO:0044598~doxorubicin metabolic process |
| GOTERM_BP_DIRECT | GO:0044597~daunorubicin metabolic process |
| GAD_DISEASE | breast cancer |
| INTERPRO | IPR020471: Aldo/keto reductase subgroup |
| INTERPRO | IPR018170: Aldo/keto reductase, conserved site |
| KEGG_PATHWAY | hsa00140: Steroid hormone biosynthesis |
| UP_SEQ_FEATURE | site: Lowers pKa of active site Tyr |
| GOTERM_BP_DIRECT | GO:0008202~steroid metabolic process |
| INTERPRO | IPR001395: Aldo/keto reductase |
| KEGG_PATHWAY | hsa05204: Chemical carcinogenesis |
| GOTERM_MF_DIRECT | GO:0047086~ketosteroid monooxygenase activity |
| GOTERM_MF_DIRECT | GO:0047718~indanol dehydrogenase activity |
| GAD_DISEASE | Adenoma\| Colorectal Neoplasms |
| PIR_SUPERFAMILY | PIRSF000097: aldo-keto reductase |
| GAD_DISEASE | bladder cancer leukemia, myeloid lung cancer |
| GOTERM_MF_DIRECT | GO:0047115~trans-1,2-dihydrobenzene-1,2-diol dehydrogenase activity |
| GOTERM_MF_DIRECT | GO:0018636~phenanthrene 9,10-monooxygenase activity |
| GOTERM_BP_DIRECT | GO:0071395~cellular response to jasmonic acid stimulus |
| GOTERM_BP_DIRECT | GO:0007584~response to nutrient |
| UP_KEYWORDS | Monooxygenase |
| UP_SEQ_FEATURE | binding site: Substrate |

| | |
|---|---|
| GAD_DISEASE | chronic obstructive pulmonary disease |
| KEGG_PATHWAY | hsa01100: Metabolic pathways |

Compared to results selected from adjusted P value, without adjusting, more detailed information present. Instead of generally affecting oxidation-reduction process, specific oxidation type can be figured out. Besides, some domain and residues related to oxidation-reduction reaction are specifically found out. Most functions are consistent with previous results. Something new is that it relates to some other metabolic process, such as doxorubicin metabolic process and daunorubicin metabolic process, it also related to cell signaling such as cellular response to jasmonic acid stimulus and response to nutrient.

**Down regulated gene functions**

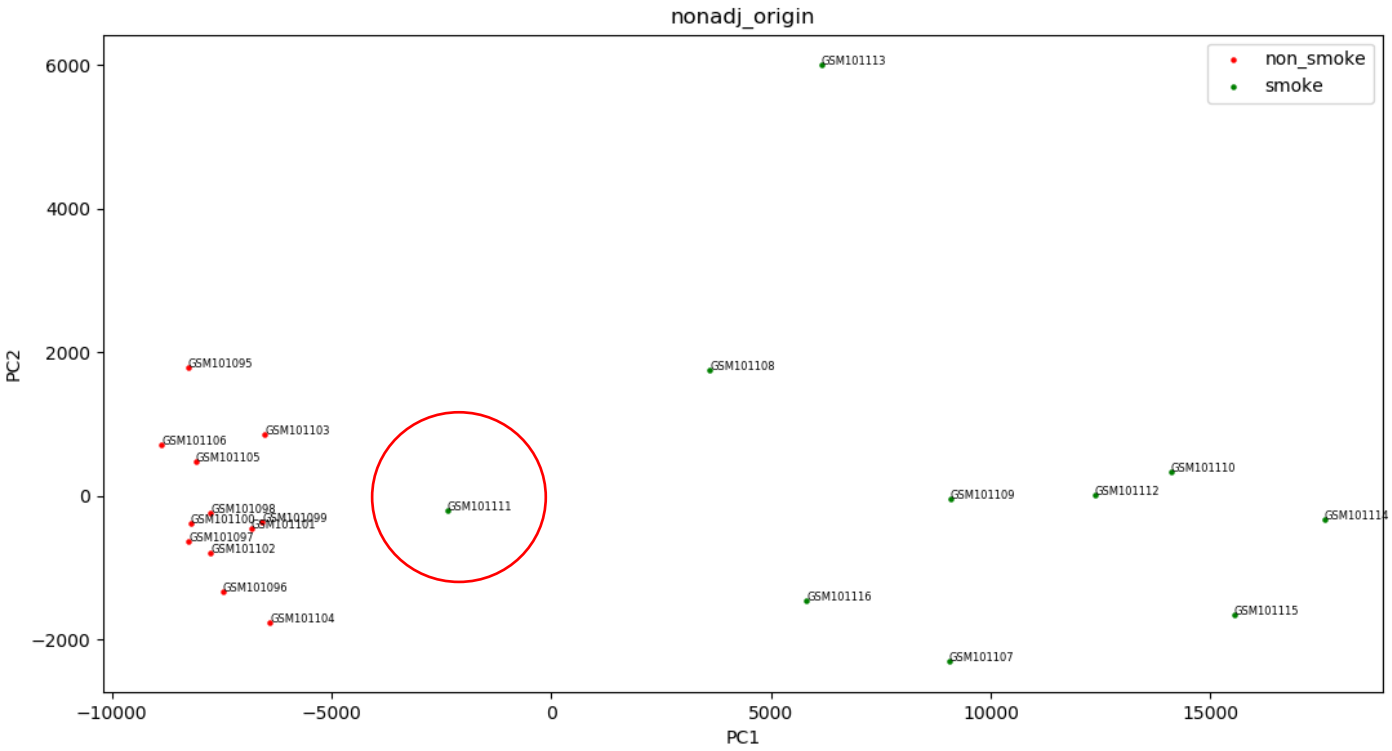
| | |
|---|---|
| GOTERM_CC_DIRECT | GO:0009986~cell surface |
| UP_SEQ_FEATURE | signal peptide |

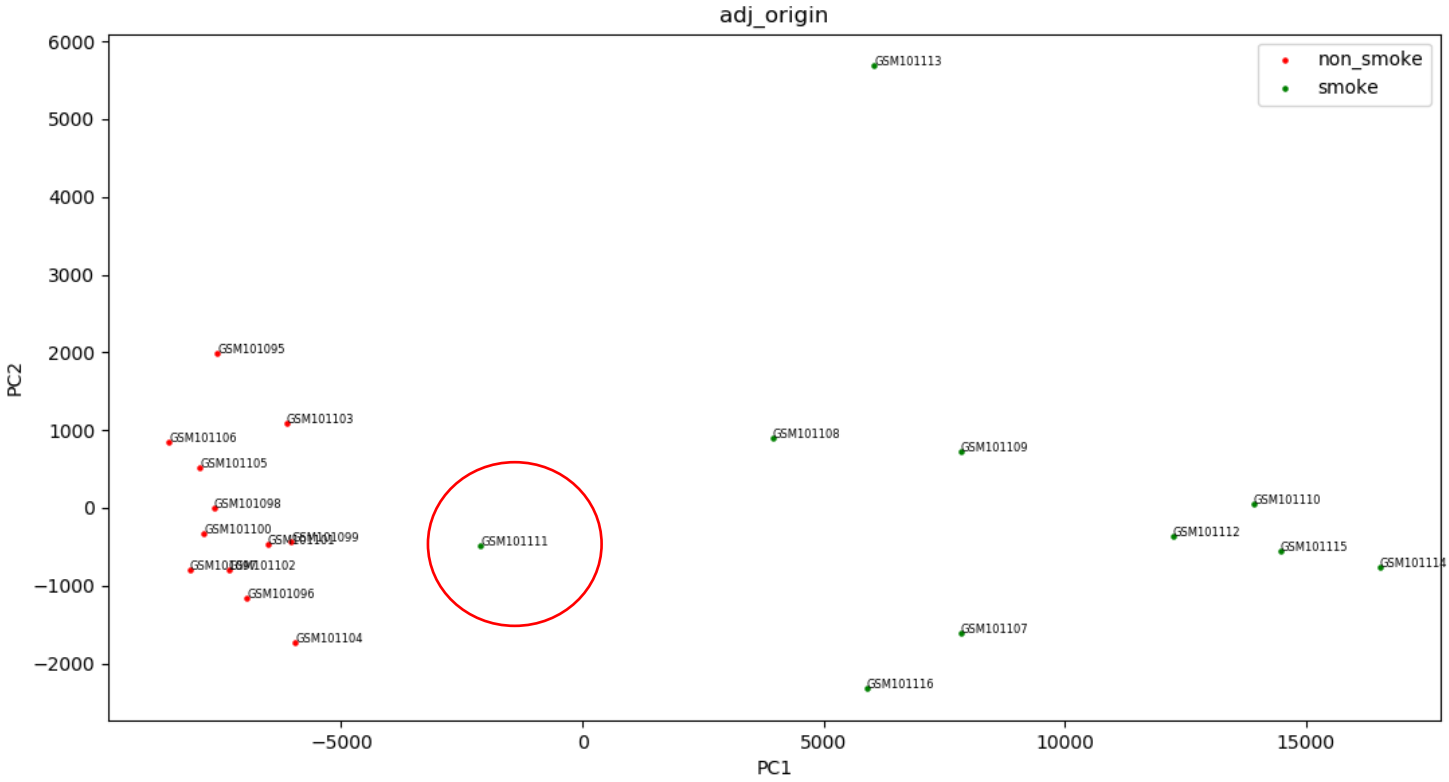
Cell surface means the external part of the cell wall and/or plasma membrane. Note that this term is intended to annotate gene products that are attached (integrated or loosely bound) to the plasma membrane or cell wall. Considering about another feature related to signal peptide, down regulated genes may relate to cell signaling, including target recognition and response.

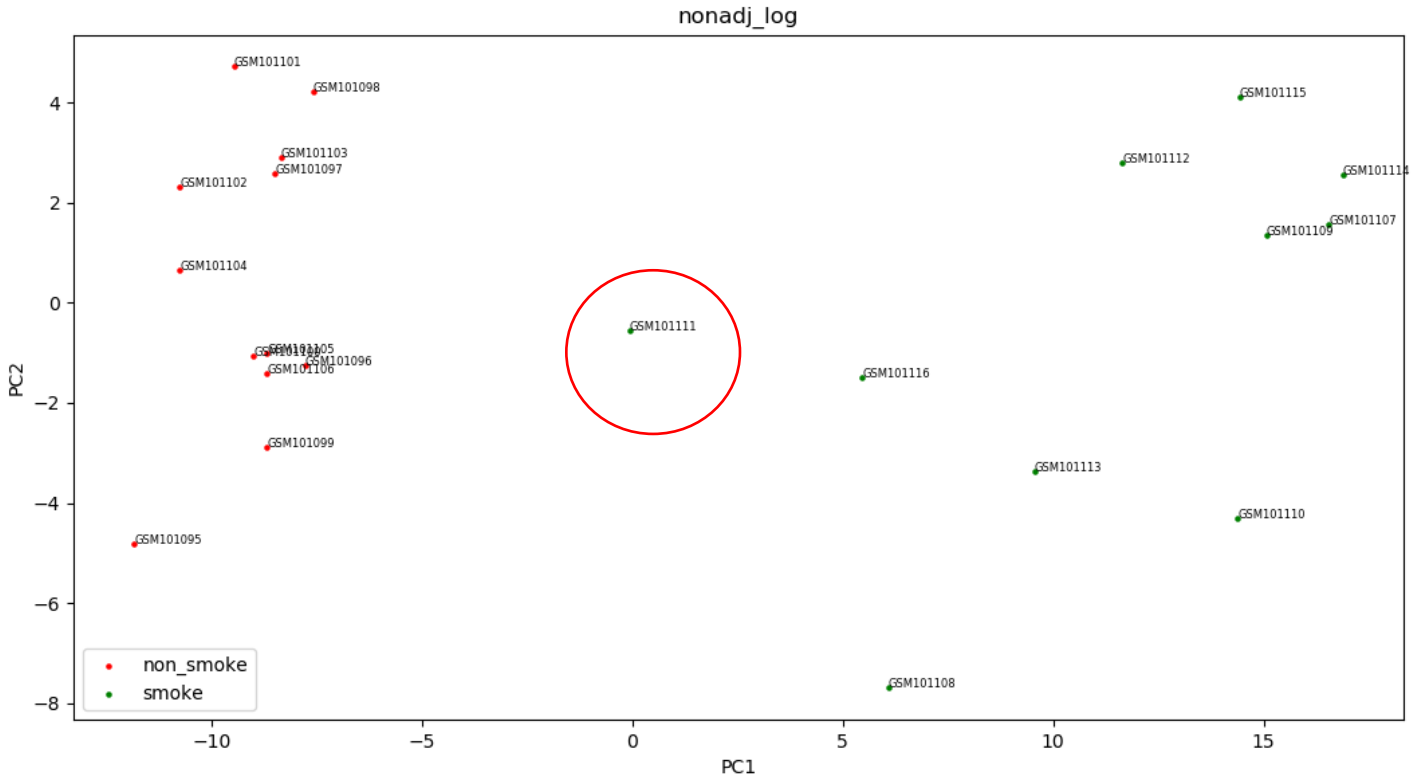
Combine results from both down regulated genes and up regulated genes, down regulated genes have less effects on samples. Cell signaling function of down regulated genes are also include in up regulated genes.

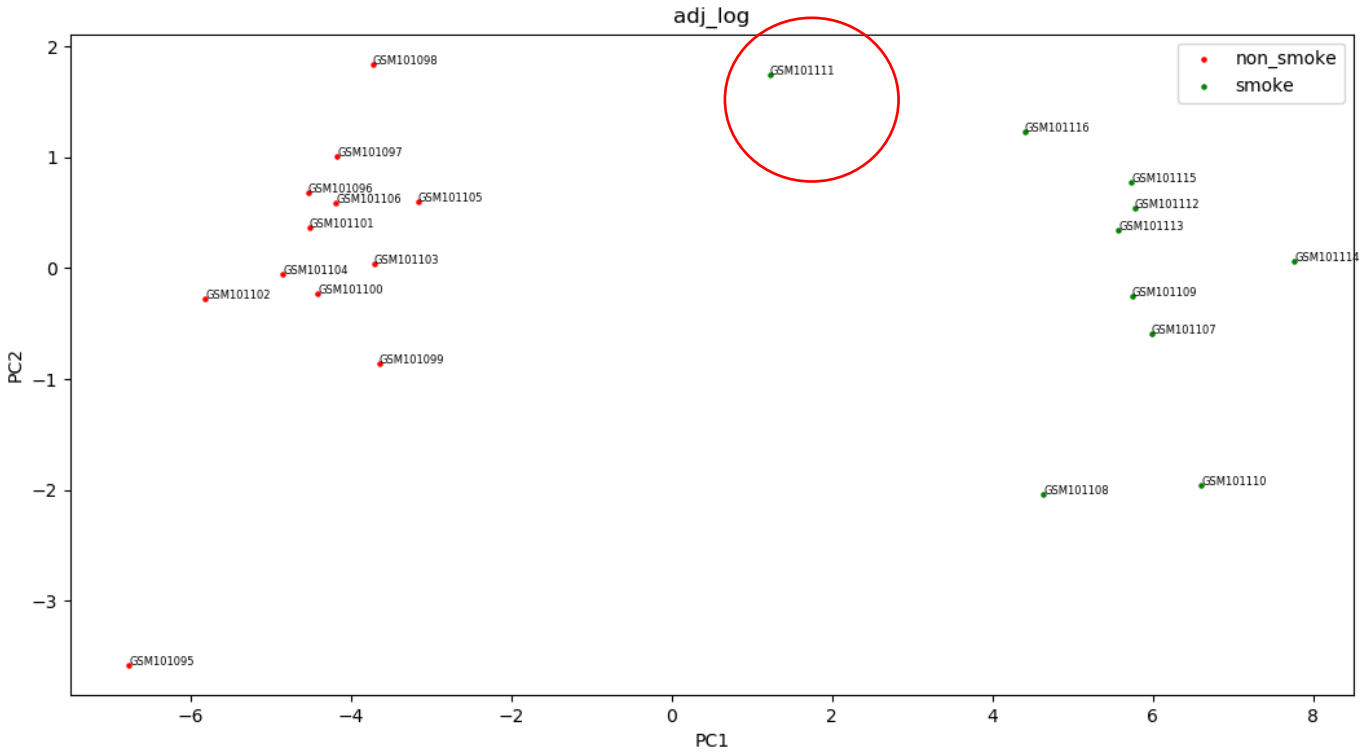
**Q3**

**principle component analysis using only DEGs**



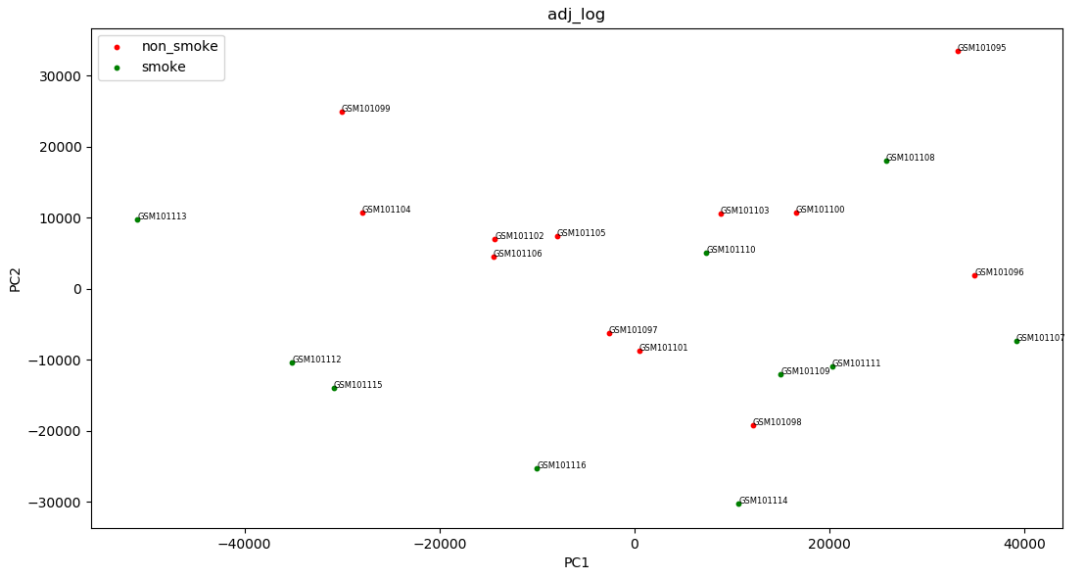



These four figures are PCA results of different combination methods of selecting DEGs that I mentioned before. The first figure use origin value and non-adjusted P to select DEGs. You can see that sample GSM10111 (in red circle) is very close to non-

smoke group. Use of adjust P value to selct DEG does not enhance clustering performance significantly as you can see in figure 2. Normalize data through doing log2 transformation helps separate GSM10111 from non-smoke group a lot. And use of both log2 normalization and adjust P value give the best clustering results. Thus, if you want to detect whether a sample belong to smokers or non-smokers, select gene features through BH adjusted P value and pre-treat data with log2 transformation may give the best prediction.

**Principle component analysis using all genes**




Conclusion from this figure is that PCA cannot separate smokers from non-smokers into two clusters based on all genes.

Finally, thanks a lot to the help of teaching assistances and my classmates!!!