NAME: K UGENDER
SR.NO: 21008
ugenderk@iisc.ac.in

Effects of Smoking

Data Analytics Assignment 3

Prog: M-tech
Dept: CSA
2023-09-18

Task

To identify the genes that respond differently to smoking in men vs women (Smoking Status x Gender vs the Smoking Status + Gender null). Using the 2-way ANOVA framework for generating the p-values for each row and visualizing P-values using histogram. Then, we estimate n_0 and False Discovery Rate (FDR) using an FDR cut-off of 0.05 to shortlist rows. By creating a short list of gene symbols from these shortlisted rows, then finding the intersection with the following gene lists: Xenobiotic Metabolism, Free Radical Response, DNA Repair, and Natural Killer Cell Cytotoxicity. Finally, reporting the intersection counts for each list, splitting them into four groups: going down in women smokers vs non-smokers and men smokers vs non-smokers, and going up in women smokers vs non-smokers and men smokers vs non-smokers.

1 Implementation

First, I loaded the Raw data set and stored the probe values for different genes. Then, I loaded the gene lists and stored them. Generated the 4×12 model matrices $M, \hat{M}$, as we have described in the class.

1.1 P-value calculation

To calculate p-values for the dataset based on a statistical analysis. Two temporary matrices are calculated using the matrices $M, \hat{M}$ as follows:

$$temp_1 = M \cdot (M^T \cdot M)^{-1} \cdot M^T$$
$$temp_2 = \hat{M} \cdot (\hat{M}^T \cdot \hat{M})^{-1} \cdot \hat{M}^T$$

Then, I computed the statistic for each data vector and the F-statistic as follows :

$$stat = \frac{i^T \cdot (a - b) \cdot i}{i^T \cdot (I - a) \cdot i}$$
$$dfn = 48 - \text{matrix\_rank}(M)$$
$$dfd = \text{matrix\_rank}(M) - \text{matrix\_rank}(\hat{M})$$

Then, the F-statistic is calculated, and the p-value using the F-distribution's cumulative distribution function as follows:

$$\text{fstat} = stat \cdot \frac{dfn}{dfd}$$
$$\text{pvalue} = \text{f.cdf}(\text{fstat}, dfd, dfn)$$

1.2 Plotting

For visualization purposes, I have plotted the histogram of p-values.

2 Results

2.1 Question 1

Used the 2-way ANOVA framework to generate p-values for each row

2.2 Question 2

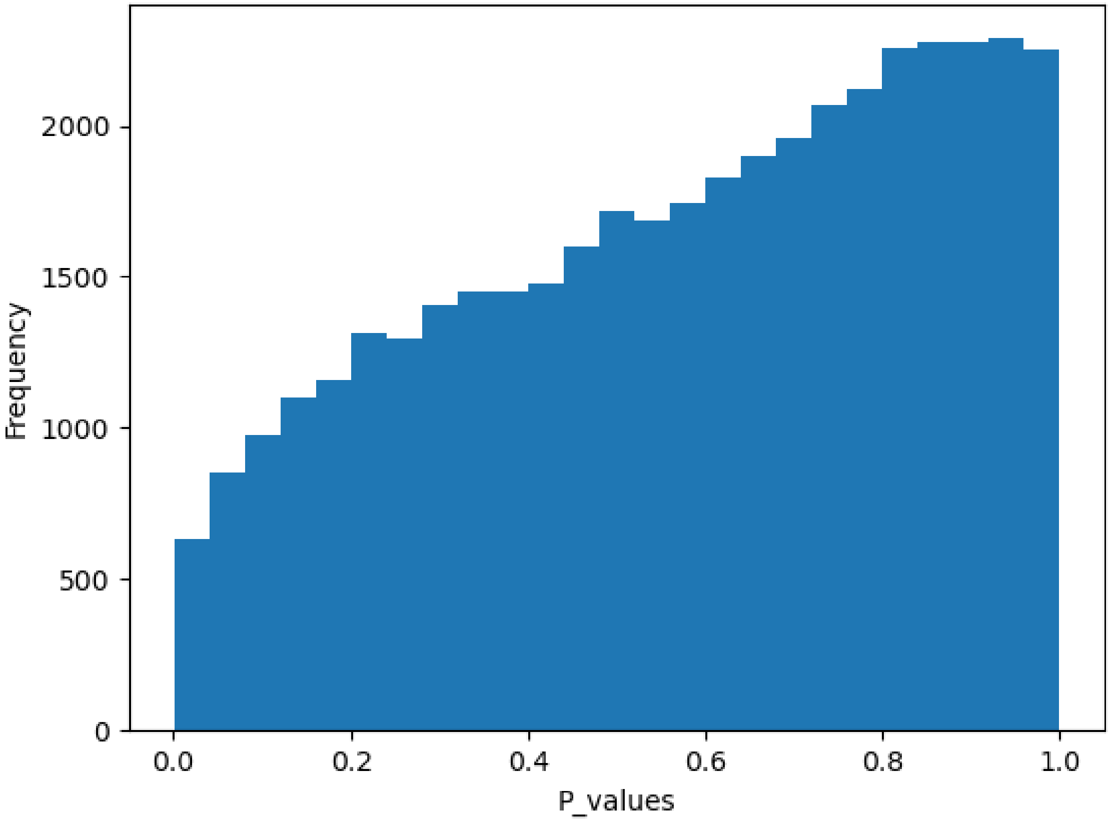

Fig. 1: Histogram of P-Values

2.3 Question 3

Based on the observed histogram, which shows a bias towards a value close to 1 (with greater density near 1 than 0), we can take a conservative estimate for n_0 as n, where n is the observed value.

2.4 Question 4

Since the estimate for n_0 is equal to n, it is not feasible to shortlist based on False Discovery Rate (FDR).

2.5 Question 5

Selected gene symbols based on their associated p-values, specifically retaining those genes with p-values less than or equal to 0.05. Then, a stored list of these significant gene symbols was compiled, pairing each symbol with its corresponding p-value.

2.6 Question 6

Intersection with Xenobiotic Metabolism Genes :

Gene Symbol	P-value
SULT1A1	0.0164
AOC2	0.0177
CYP2S1	0.0100
AADAC	0.0492
HNF4A	0.0367
AS3MT	0.0105

Table- 1: Xenobiotic Metabolism Genes and P-values

Intersection with Free Radical Response Genes :

No gene found in the intersection

Table- 2: Free Radical Response Genes and P-values

Intersection with DNA Repair Gene :

Gene Symbol	P-value
PNKP	0.0490

Table- 3: DNA Repair Genes and P-value

Intersection with Natural Killer Cell Cytotoxicity Genes :

Gene Symbol	P-value
IFNG	0.0424
KLRC2	0.0187
PTPN6	0.0087
HLA-C	0.0243
PRF1	0.0475
HLA-E	0.0391
HLA-G	0.0207

Table- 4: Natural Killer Cell Cytotoxicity Genes and P-values

2.7 Question 7

Intersection counts for each list, spliting into four groups: going down in women smokers vs non-smokers, men smokers vs men non-smokers and going up in women smokers vs non-smokers, men smokers vs men non-smokers.
For Xenobiotic metabolism Genes:

Comparison	Upregulated Genes	Downregulated Genes
Smokers vs. Non-Smokers (Women)	SULT1A1, AOC2, CYP2S1, HNF4A	AADAC, AS3MT
Smokers vs. Non-Smokers (Men)	AADAC, HNF4A, AS3MT	SULT1A1, AOC2, CYP2S1, HNF4A

Table- 5: Xenobiotic metabolism Genes

For DNA Repair Genes:

Comparison	Upregulated Genes	Downregulated Genes
Smokers vs. Non-Smokers (Women)	PNKP	-
Smokers vs. Non-Smokers (Men)	-	PNKP

Table- 6: DNA Repair Genes

For Natural Killer Cell Cytotoxicity Genes:

Comparison	Upregulated Genes	Downregulated Genes
Smokers vs. Non-Smokers (Women)	PTPN6, HLA-C, HLA-E, HLA-G	IFNG, KLRC2, PRF1
Smokers vs. Non-Smokers (Men)	IFNG, KLRC2, PRF1, HLA-E, HLA-G	PTPN6, HLA-C

Table- 7: Natural Killer Cell Cytotoxicity Genes

For OverAll types of Genes :

Comparison	Upregulated Genes	Downregulated Genes
Smokers vs. Non-Smokers (Women)	SULT1A1, AOC2, CYP2S1,	AADAC, AS3MT,
	PNKP, PTPN6, HLA-C,	KLRC2, PRF1
	HLA-E, HLA-G,HNF4A,	IFNG,
Smokers vs. Non-Smokers (Men)	AADAC, HNF4A, AS3MT,	SULT1A1, AOC2, CYP2S1,
	KLRC2, PRF1, HLA-E,	PTPN6, HLA-C, HLA-E,
	IFNG, HLA-G	HLA-G, HNF4A, PNKP,

Table- 8: Overall Differential Gene Expression in Smokers vs. Non-Smokers