Data Analytics - Assignment IV

Anup Patel (Sr.No. - 15474)
M.tech CSA

October 30, 2019

Effect of Smoking

Part 1

To identify genes which respond differently to smoke in men vs. women (Smoking Status X Gender model vs. Smoking Status + Gender null model)
We computed A and A' using :

$$h = AB + Error$$

h: gene expression
A: Alternative Hypothesis
A': Null Hypothesis
B: mean vector

For Null Model A' :

$$\begin{bmatrix} h_1 \\ h_2 \\ . \\ . \\ . \\ . \\ h_{48} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ . & & & \\ . & & & \\ . & & & \\ . & & & \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} male \\ female \\ nonsmoker \\ smoker \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ . \\ . \\ \epsilon_{48} \end{bmatrix}$$

For Alternative Model A :

$$\begin{bmatrix} h_1 \\ h_2 \\ . \\ . \\ . \\ . \\ h_{48} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ . & & & \\ . & & & \\ . & & & \\ . & & & \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} male_{nonsmoker} \\ male_{smoker} \\ female_{nonsmoker} \\ female_{smoker} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ . \\ . \\ \epsilon_{48} \end{bmatrix}$$

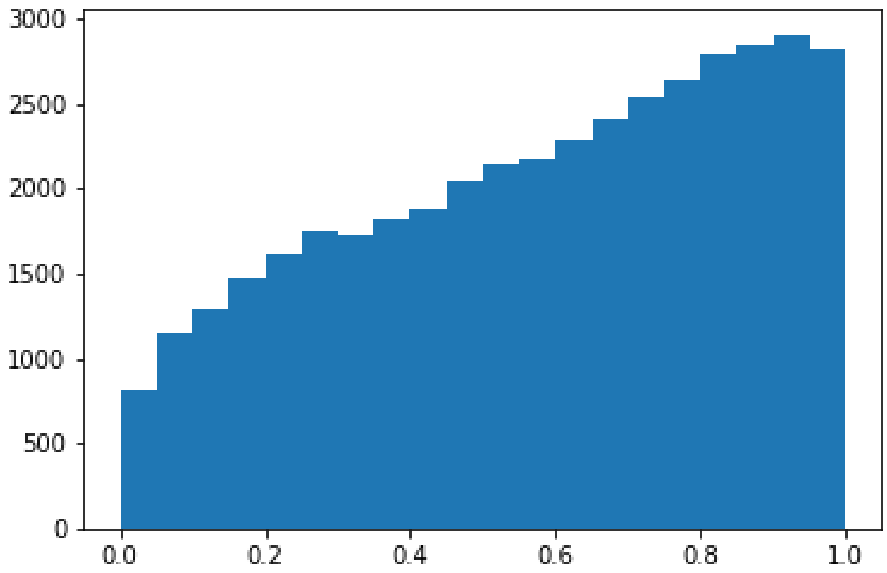
I had computed F-Statistic by using formula:

$$\text{F-statistic } \hat{f} = \frac{\vec{\hat{h}}^T (A(A^T A)^\dagger A^T - A'(A'^T A')^\dagger A'^T)\vec{\hat{h}}}{\vec{\hat{h}}^T (I-(A(A^T A)^\dagger A^T)\vec{\hat{h}}} * \frac{n-rank(A)}{rank(A)-rank(A')}$$

After this i had computed p-value by using scipy library (refer code part)

Part 2

To draw the histogram of p-values



Part 3

Refer code

Part 4

Refer code

Part 5

Refer genes-symbol-list.txt file

Part 6

Genes Intersection with Xenobiotic metabolism ::
['SULT1A1', 'AOC2', 'CYP2S1', 'AADAC', 'HNF4A', 'AS3MT']

Genes Intersection with Free Radical Response :: None

Genes Intersection with DNA Repair ::
[PNKP]

Genes Intersection with Natural Killer Cell Cytotoxicity ::
['IFNG', 'KLRC2', 'PTPN6', 'HLA-C', 'PRF1', 'HLA-E', 'HLA-G']

Part 7

Intersection Count:

Xenobiotic metabolism :: 6
Free Radical Response :: 0
DNA Repair :: 1
Natural Killer Cell Cytotoxicity :: 7

Groups are as follows:

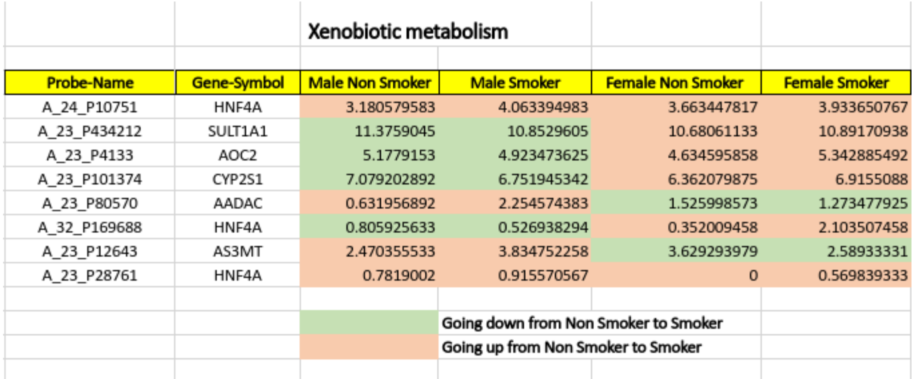
Female Smokers up genes ::
['HLA-C', 'CYP2S1', 'PNKP', 'HLA-E', 'SULT1A1', 'AOC2', 'HLA-G', 'HNF4A', 'PTPN6']

Female Smokers Down genes ::
['KLRC2', 'IFNG', 'AADAC', 'PRF1', 'AS3MT']

Male Smokers up genes::
['KLRC2', 'IFNG', 'AADAC', 'HLA-G', 'PRF1', 'AS3MT', 'HNF4A', 'HLA-E']

Male Smokers down genes::
['HLA-C', 'CYP2S1', 'PNKP', 'HLA-E', 'SULT1A1', 'AOC2', 'HLA-G', 'HNF4A', 'PTPN6']

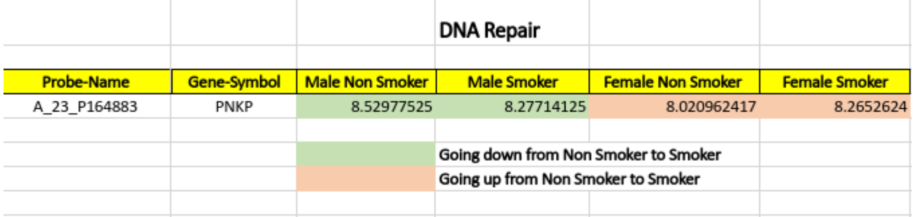
Here is the detailed table of how i had splitted in 4 Groups :



Xenobiotic metabolism

Probe-Name	Gene-Symbol	Male Non Smoker	Male Smoker	Female Non Smoker	Female Smoker
A_24_P10751	HNF4A	3.180579583	4.063394983	3.663447817	3.933650767
A_23_P434212	SULT1A1	11.3759045	10.8529605	10.68061133	10.89170938
A_23_P4133	AOC2	5.1779153	4.923473625	4.634595858	5.342885492
A_23_P101374	CYP2S1	7.079202892	6.751945342	6.362079875	6.9155088
A_23_P80570	AADAC	0.631956892	2.254574383	1.525998573	1.273477925
A_32_P169688	HNF4A	0.805925633	0.526938294	0.352009458	2.103507458
A_23_P12643	AS3MT	2.470355533	3.834752258	3.629293979	2.58933331
A_23_P28761	HNF4A	0.7819002	0.915570567	0	0.569839333

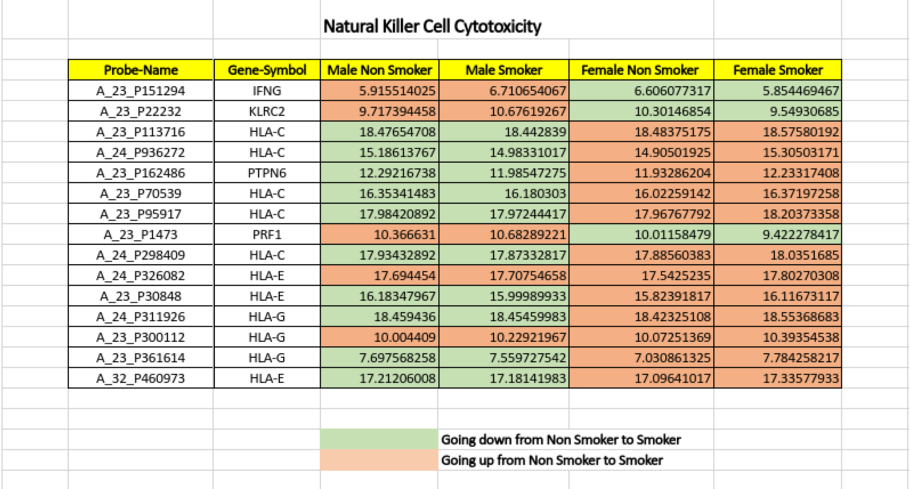
🟩	Going down from Non Smoker to Smoker
🟧	Going up from Non Smoker to Smoker



DNA Repair

Probe-Name	Gene-Symbol	Male Non Smoker	Male Smoker	Female Non Smoker	Female Smoker
A_23_P164883	PNKP	8.52977525	8.27714125	8.020962417	8.2652624

🟩	Going down from Non Smoker to Smoker
🟧	Going up from Non Smoker to Smoker



Natural Killer Cell Cytotoxicity

Probe-Name	Gene-Symbol	Male Non Smoker	Male Smoker	Female Non Smoker	Female Smoker
A_23_P151294	IFNG	5.915514025	6.710654067	6.606077317	5.854469467
A_23_P22232	KLRC2	9.717394458	10.67619267	10.30146854	9.54930685
A_23_P113716	HLA-C	18.47654708	18.442839	18.48375175	18.57580192
A_24_P936272	HLA-C	15.18613767	14.98331017	14.90501925	15.30503171
A_23_P162486	PTPN6	12.29216738	11.98547275	11.93286204	12.23317408
A_23_P70539	HLA-C	16.35341483	16.180303	16.02259142	16.37197258
A_23_P95917	HLA-C	17.98420892	17.97244417	17.96767792	18.20373358
A_23_P1473	PRF1	10.366631	10.68289221	10.01158479	9.422278417
A_24_P298409	HLA-C	17.93432892	17.87332817	17.88560383	18.0351685
A_24_P326082	HLA-E	17.694454	17.70754658	17.5425235	17.80270308
A_23_P30848	HLA-E	16.18347967	15.99989933	15.82391817	16.11673117
A_24_P311926	HLA-G	18.459436	18.45459983	18.42325108	18.55368683
A_23_P300112	HLA-G	10.004409	10.22921967	10.07251369	10.39354538
A_23_P361614	HLA-G	7.697568258	7.559727542	7.030861325	7.784258217
A_32_P460973	HLA-E	17.21206008	17.18141983	17.09641017	17.33577933

🟩	Going down from Non Smoker to Smoker
🟧	Going up from Non Smoker to Smoker